Deliverable D6.3 – Final Use Case Evaluation Report

OPENing UP new methods, indicators and tools for peer review, impact measurement and dissemination of research results

Project acronym: OpenUP
Grant Agreement no. 710722

<table>
<thead>
<tr>
<th>Deliverable/Milestone information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deliverable number and name</td>
</tr>
<tr>
<td>Due date</td>
</tr>
<tr>
<td>Delivery</td>
</tr>
<tr>
<td>Work Package</td>
</tr>
<tr>
<td>Lead Partner for deliverable</td>
</tr>
<tr>
<td>Author</td>
</tr>
<tr>
<td>Reviewers</td>
</tr>
<tr>
<td>Approved by</td>
</tr>
<tr>
<td>Dissemination level</td>
</tr>
<tr>
<td>Version</td>
</tr>
</tbody>
</table>
Table 1. Document revision history

<table>
<thead>
<tr>
<th>Issue Date</th>
<th>Version</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018-04-25</td>
<td>0.1</td>
<td>Outline</td>
</tr>
<tr>
<td>2018-07-23</td>
<td>0.2</td>
<td>Contributions from Pilot Leads</td>
</tr>
<tr>
<td>2018-07-31</td>
<td>0.3</td>
<td>Incomplete draft for first review by Vilte Banelyte</td>
</tr>
<tr>
<td>2018-08-24</td>
<td>0.4</td>
<td>Complete draft including all pilot study reports &amp; draft conclusions (to be discussed with WP6 team)</td>
</tr>
<tr>
<td>2018-08-31</td>
<td>0.5</td>
<td>Final draft for internal review</td>
</tr>
<tr>
<td>2018-09-13</td>
<td>1.0</td>
<td>Final version addressing comments and amendments by Vilte Banelyte and Saskia Woutersen</td>
</tr>
</tbody>
</table>
Disclaimer

Unless otherwise noted, this work is licensed under the Creative Commons Attribution 4.0 International License. [http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/). In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of OpenUP consortium and can in no way be taken to reflect the views of the European Union.

OpenUP is a project partially funded by the European Union

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 28 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors. ([http://europa.eu.int/](http://europa.eu.int/))

Unless otherwise noted, this work is licensed under the Creative Commons Attribution 4.0 International License. [http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/)
# Table of Contents

Disclaimer ............................................................................................................................................ 2
Table of Contents ................................................................................................................................ 3
Summary .............................................................................................................................................. 6

1. **Executive Summary: Synthesis and Lessons Learned** ..................................................................... 7
1.1. Open Peer Review (OPR) ................................................................................................................... 8
   1.1.1. OPR at conferences ..................................................................................................................... 8
   1.1.2. OPR of research data .................................................................................................................. 8
1.2. Open Research Data: requirements and preconditions ..................................................................... 9
1.3. Reaching and involving stakeholders outside of academia ................................................................. 10
1.4. Impact Assessment and Open Science ............................................................................................. 10
   1.4.1. Metrics to assess impact of dissemination activities directed to specific target groups .......... 10
   1.4.2. Metrics to incentivise open research data (in biomedicine and beyond) ................................. 10
   1.4.3. Implications of academic piracy for impact assessment ............................................................. 11
1.5. Policy and action recommendations ................................................................................................. 11
   1.5.1. Towards a sustainable open science environment .................................................................. 11
   1.5.2. Developing scholarly publication models ............................................................................... 11

2. **OpenUP Pilots Goals & Evaluation Approach** ............................................................................... 12

3. **Final Evaluation Reports** ............................................................................................................... 13
   3.1. **Pilot 1: Open Peer Review for Conferences** ........................................................................... 13
      3.1.1. Introduction ............................................................................................................................... 13
      3.1.2. Venue 1: EMVA 2017 ............................................................................................................. 14
      3.1.3. Venue 2: eHealth2018 Master Student Competition ................................................................. 14
      3.1.4. Results .................................................................................................................................... 15
         3.1.4.1. Open Source Conference Management Software Repository ....................................... 15
         3.1.4.2. EMVA Feedback Survey .................................................................................................... 15
         3.1.4.3. eHealth2018 Student Competition Feedback Survey ......................................................... 15
         3.1.4.4. Presentations and discussions at conferences ................................................................. 16
      3.1.5. Lessons learned ....................................................................................................................... 16
   3.2. **Pilot 2: Open Peer Review for Research Data in Social Sciences** .......................................... 16
      3.2.1. Introduction ............................................................................................................................. 16
      3.2.2. Interviews ............................................................................................................................... 17
      3.2.3. Origin, motivations and organisational features ................................................................. 18
      3.2.4. Goals and main features of the database .......................................................................... 18
      3.2.5. Data quality assessment process ............................................................................................ 19
      3.2.6. Opinion on open access of data and peer review ................................................................. 20
      3.2.7. Survey of the HDM users ...................................................................................................... 20
      3.2.8. Results .................................................................................................................................... 21
         3.2.8.1. Respondent’s profile ....................................................................................................... 21
Summary

The seven pilot studies run under the sixth Work Package of the OpenUP project were successful and collected valuable input from the communities involved. Key stakeholders involved in the pilots were researchers, publishers, data providers, institutions, research projects, and general public stakeholders. The pilots contributed to raising awareness and increasing skills related to the tested open science approaches among the involved communities, and generated lessons learned and an evidence base on various aspects of the tested approaches. Lessons learned and insights about applicability of the tested approaches to a specific field and community as well as their impact on distinct research communities and researchers are gathered in this report.

The synthesis and lessons learned from all pilots are assembled in the Executive Summary. The detailed pilot evaluation reports are included in Section 3.

Key findings and lessons learned from the pilots include:

- **Open Peer Review:**
  - Open Peer Review for conferences was well received by the involved communities
  - Open Peer Review for data supports Open Research Data
  - Principles of Open Peer Review of research data

- **Open Research Data:**
  - Publishing, managing and reviewing research data: Recognition of and rewards for activities are needed
  - Diversity of research data: Definition suitable for the Humanities needed
  - Principles of research data management and data journals in the Humanities
  - Fostering data publication and quality assessment: Principles for creating an Open Research Database
  - Motivations for sharing research data in Social Sciences and the Humanities
  - Barriers to open research data (focus: Humanities and Life Sciences/Biomedicine)

- **Reaching and involving stakeholders outside of academia:**
  - Transferring the research workflow to the web involving the general public works
  - Diverse approach for defining and supporting open science dissemination is needed

- **Impact assessment:**
  - Using altmetrics for assessing the impact of research dissemination at specific target groups is not straightforward
  - Suggested measures for establishing open research data incentives
  - Use of pirated academic publications skew demand for key library resources

- **Policy and action recommendations:**
  - Strong commitment by organisations, funders, and policy makers needed
  - Support and training for researchers needed
  - Encourage Open Online Research in research projects
  - Lessons learned from academic piracy for shaping business models in scholarly publishing
1. Executive Summary: Synthesis and Lessons Learned

The goal of OpenUP’s 6th work package was to implement, test, and verify the outputs and frameworks from WP3-WP5. To that end, during the course of 2017-2018, seven pilot studies were run in specific settings and in close collaboration with various research communities. The seven pilots are attributed to three Use Cases, which correspond to the OpenUP pillars (see Table 1).

Table 1. OpenUP Use Cases and Pilots

<table>
<thead>
<tr>
<th>Innovative peer review applied to specific contexts</th>
<th>Innovative dissemination of research output in specific contexts</th>
<th>Measuring impact of research output applied to specific contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot 1: Open Peer Review for Conferences</td>
<td>Pilot 4: Transferring the research lifecycle to the web (Open Science Repositories)</td>
<td>Pilot 6: Reflexivity of metrics on medical research and dissemination practices</td>
</tr>
<tr>
<td><strong>Goal:</strong> Test an open peer review workflow in a conference setting</td>
<td><strong>Goal:</strong> Investigate whether qualitative research, in particular data analysis and data collection, can be transferred to open online groups</td>
<td><strong>Goal:</strong> Engage with and reflect on current practices and stances of handling data in the biomedical research enterprise, in order to identify barriers and enablers for Open Data, and to derive recommendations for their incentivisation (or possible metrification).</td>
</tr>
<tr>
<td>Pilot 2: Open Peer Review for Research Data</td>
<td>Pilot 5: Addressing &amp; reaching businesses and the public with research output</td>
<td>Pilot 7: Piratical demand as a form of impact indicator and reaching unexpected audiences</td>
</tr>
<tr>
<td><strong>Goal:</strong> Investigate the applicability of (open) peer review to research data in disciplines related to Social Sciences</td>
<td><strong>Goal:</strong> Analyse and test how disseminated research results from the Energy area can be made more interesting, appealing, and usable for target audiences beyond the research community</td>
<td><strong>Goal:</strong> Conduct a quantitative, statistical and econometric analysis of large scale datasets on the supply of and demand for scholarly works on various illegal platforms</td>
</tr>
<tr>
<td>Pilot 3: A data journal for the Arts and Humanities</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Goal:</strong> Define a framework for a data journal in the Humanities and provide a related action plan</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All pilot studies achieved the projected interim results after the first implementation phase and collected valuable input from the communities involved. Key stakeholders involved in the pilots were researchers, publishers, data providers, institutions, research projects, and general public stakeholders. All pilots run under OpenUP contributed to

- Raising awareness and increasing skills related to the tested open science approaches among the involved communities.
- Generating lessons learned and an evidence base on various aspects of the tested approaches, in particular their applicability to a specific field and contacts, and what impacts they have in distinct research communities and researchers (considering their gender, career stage, country, ethnicity).

The pilots successfully connected and collaborated with communities who were particularly interested in applying open science approaches and providing alternative access to scientific output. What OpenUP did was, on the one hand, raise their interest and support them in achieving their open science goals. On

the other hand, we had intensive dialogues with them, collected their problems and challenges that they encountered while putting the tested approaches into practice. Beyond testing the selected open science approaches in particular application contexts, we sustainably supported the involved communities to adopt the tested practices and solutions beyond the duration of OpenUP. In two cases, the collaboration activities with the communities resulted in ulterior funding and/or long-term cooperation between the pilot leading institutions and the community contacts (Pilot 4, Pilot 6).

During the final evaluation phase, all pilots were able to achieve the main goals previously defined:

- The two design studies (Pilot 2, Pilot 3) provided the results from the analysis of current dataset management and sharing practices in the Social Sciences, and on research data management in the Humanities. Both pilots provided a qualitative analysis of the interviews and survey on the applicability of (open) peer review to datasets in Social Sciences, and of the state-of-the-art report and requirements analysis for data journals in the Humanities.
- Pilots producing software (Pilot 1, Pilot 4) delivered testable and tested prototypes of adapted CMS software and an open online research tool. The pilots delivered a qualitative and quantitative analysis of the feedback from the participants testing the CRM tool, and of the test round of the OpenOnlineResearch tool.
- Pilots 5 and 6 delivered draft recommendations and guidelines for reaching stakeholders from the business sector and the general public; and draft community profiles of Translational Medical Research communities. Both pilots delivered a qualitative and quantitative analysis of the resulting dissemination material and impact, and of the Open Data related Altmetrics usage in translational research.
- Finally, the statistical analysis pilot (Pilot 7) delivered a report of the piratical access provider dataset enrichment, and a first visualisation of the data created with the developed online service allowing real-time exploration of the dataset. The pilot delivered a micro- and macroeconomic model of country level usage of piratical scholarly publications.

The following paragraphs summarise key findings and lessons learned from all pilots related to the main challenges that the communities are currently facing. These findings provide additional evidence for successes achieved through open science as well as persisting issues that need to be addressed further.

1.1. Open Peer Review (OPR)

1.1.1. OPR at conferences

- OPR for conferences was well received by the involved communities

   Overall, the communities participating in Pilot 1 expressed a strong acceptance of the tested OPR process and would support it again. The conference participants’ greatest fears associated with OPR included: biased/whitewashed reviews due to non-anonymity; backlash for bad reviewing (e.g. over other channels/private email); and added effort and risk for reviews outside one’s own expertise (layman reviews)

1.1.2. OPR of research data

- OPR for data supports Open Research Data

   The 2nd OpenUP pilot study showed that one of the main successful features of the Human Mortality Database (HMD) is the platform’s transparent way of data managing and sharing. In particular this concerns the two central data validation phases, which adhere to OPR principles. This shows that adding OPR functionalities gives an added value to research communities, and that it substantially supports reproducibility and quality in science.

- Principles of OPR of research data

   There are similarities between the peer review of articles and peer review of data. However, data review includes more issues to manage. Research datasets can be complex, multifaceted, and dynamic information objects. Therefore, peer review of data can be a cumbersome task. Review of data should
involve the checking and evaluation of the methodology of the data collection, and the examination of the software code used to process the data.

- **Publishing, managing and reviewing research data: Recognition of and rewards for activities are needed**

Interviews with the HMD community showed that publications of scientific papers are generally considered more important than managing a database. Analysis quality check of research data to be provided is a researcher activity in itself and should be valued as such. We agree with the Next Generation Metrics EG report and the OSPP recommendations with regards of expanding evaluation of research to other than just publication outputs.

1.2. Open Research Data: requirements and preconditions

- **Diversity of research data: Definition suitable for the Humanities needed**

Research data is as diverse as the disciplines that use them. In the on-going European discourse around open data, prevailing definitions and understandings of data relate to measurable, collected, reported, and analysed data. This is not a common concept in the Humanities, where content and research cannot be related to numbers and statistics.

To address research communities in the Humanities, an adapted definition of data is needed. We suggest defining research data in the Humanities as the evidence used to inform or support research conclusions.

- **Principles of research data management and data journals in the Humanities**

Due to the varied fields of study involved in the Humanities, contextualization of data is needed to understand research data management. A data journal usually provides templates for data description and offers researchers guidance on where to deposit and how to describe and present their data. A data journal framework should include assignment of persistent identifiers (PIIDs) to datasets, peer review of data, metadata information and technical check, links to related outputs (journal articles), facilitation of data citation, standards compliance, and discoverability (indexing of the data).

- **Fostering data publication and quality assessment: Principles for creating an Open Research Database**

As the HMD use case shows, Social Sciences data users appreciate that the platform makes source data available, provides transparent data processing procedures, detailed documentation, and that it is trustworthy. These are also important features for any other research data service or platform. Guiding principles for creating this open research data platform were comparability, flexibility, accessibility and reproducibility. These principles are applicable to other research areas as well.

- **Motivations for sharing research data in Social Sciences and the Humanities**

A campus-wide survey conducted in 2015 at the University of Lille suggests that students in Social Sciences the Humanities are more interested in achieving impact and visibility (international repository) and disciplinary specificity (laboratory) rather than in creating a multidisciplinary, national or institutional solution.

- **Barriers to open research data (focus: Humanities and Life Sciences/Biomedicine)**

Pilot 3 showed that one of the main obstacles of data sharing in the Humanities is the generally closed world of scientific discourse. The dispersed research communities often fail to connect to one another because of the language barriers. Humanities scholars very often publish in their national languages, and the trend is to continue doing so in the future. Europe lacks an integrated database of published journals in various national languages. A database of this kind could be a sort of 'who's who' within a particular field of research.

Pilot 6 showed that there is a lack of information and oversight in the biomedical field. This is due to missing infrastructures as well as heterogeneous and non-standardized publication practices regarding the provision of open data. Even if there are guidelines such as FAIR data, useful criteria for “good data” which account for the heterogeneity of research fields such as biomedicine are still missing. To bring this to success we need to consider field specific solutions or standards. Another reason is the lack of training and technical expertise for opening and curating the data.
Similarly, a lack of standards and common guidelines in data management has been observed in the Humanities (Pilot 3). Due to this, it is very difficult to connect data (Humanities). Lack of standardized workflows for data curation, sharing and publishing is, in fact, a barrier in many disciplines.

1.3. Reaching and involving stakeholders outside of academia

- **Transferring the research workflow to the web involving the general public works**
The results of the test rounds done in context of Pilot 4 confirmed that online applications such as the OpenOnlineResearch (OOR) tool can enable citizens to gather and analyse data online and openly. By means of a model investing in social moderation, we demonstrated that open online interpretation of qualitative data is feasible and that yet unused parts of the research cycle can be opened to wider ranges of collaborators both within and outside academia.

We learned that a simple tool such as OOR works without the need for having detailed instructions for the functionalities of the tool. Input of scientists, however, is essential for the formulation of sound research questions and instructions. Online collaboration needs moderation (either technically or by humans) to settle differences. However, we found that conflicts were rare and that participants were willing to collaborate in most cases.

- **Diverse approach for defining and supporting open science dissemination is needed**
There is no one size fits all. We came up with guidelines to support communication to two large target groups: businesses and the general public. The test run done in context of Pilot 5 showed that the guidelines were very useful for improving the dissemination strategy of the project. However, the division in the two above mentioned target groups has proven not to be necessarily evident or useful for all projects in the Energy area. The division into these two main target groups is rather arbitrary and cannot necessarily be applied to other specific target groups. A more diverse approach to defining and supporting open science dissemination is needed (e.g. including other target groups or addressing trans-disciplinary questions). The guidelines that we authored already provide good guidance; however only for planning a dissemination strategy tailored to these two target audiences, and not beyond. For future research it would be relevant to explore other ways to structure the guidelines and their content to provide additional guidance for the points that our guidelines fail to provide substantial support.

1.4. Impact Assessment and Open Science

1.4.1. Metrics to assess impact of dissemination activities directed to specific target groups

- **Using altmetrics for assessing the impact of research dissemination at specific target groups is not straightforward**
At the example of Twitter, in Pilot 5 we tried to test if alternative metrics such as re-tweets and likes from Twitter users can be used as a meaningful indicator for assessing impact in specific stakeholder groups. Our results suggest that it is not possible to clearly establish to which target group the reached individuals belong. Even if the account information clearly include references to interests in research topics it is difficult to draw conclusions about the stakeholder or target group an individual belongs to (e.g. the individual accounts can be used for various purposes or do not disclose enough information about which target group the user belongs to). Therefore, it is questionable if the resulting analysis can be used as a meaningful indicator for assessing a dissemination impact of a research project in specific stakeholder groups.

1.4.2. Metrics to incentivise open research data (in biomedicine and beyond)

- **Suggested measures for establishing open research data incentives**
To establish metrics to incentivize Open Data, research managers should first of all provide services on how to open data which are field specific, and which refer to the regulations in the biomedical field. There should be more institutional and organizational support for using services which make data sets
citable or trackable. Lack of institutional visibility is due to a shortcoming in regarding the way data are disseminated and made visible.

1.4.3. Implications of academic piracy for impact assessment

- **Use of pirated academic publications skew demand for key library resources**
  As it stands now, academic piracy seems to be an unstoppable force. In our 7th pilot study we showed that shadow libraries are now an integral part of the systems of scholarly communication. They are part of the everyday routine of scholars of both the developed and the developing countries. The trend is showing a radical shift towards data-centred business models. In the last decades academia underwent a substantial degree of quantification, where not just citations, but all other aspects of scholarly work became measurable and consequently thoroughly measured. However, there is no reliable, systematic insight into the use of the resources provided by piracy bays for scholarly publications. Only thanks to the dataset provided by one of the shadow library communities, we were able to analyse it for the years 2012-2015, which delivered an insight to these use and demand indicators that usually are under the radar. Since the official usage statistics of resources available through shadow libraries are not included in the statistics determining the demand for key library resources, the same demand remains under-reported.

1.5. Policy and action recommendations

1.5.1. Towards a sustainable open science environment

- **Strong commitment by organisations, funders, and policy makers needed**
  Pilot 2 showed that having a strong commitment of the organisation in supporting development of data infrastructure is crucial. In general, unfortunately, this is not always the case and not only support by organisations (institutional policies & funding of activities) is lacking, but also sustainable funding models. An example of institutional support is when researchers are formally endorsed to spend half of their work time on research activities related to the analysis and curation of data before they are made publicly available.

- **Support and training for researchers needed**
  Pilot 2 showed that long-term financial support (beyond the project duration), institutional policy endorsement on open data as well as a formal recognition of scientists for the efforts in data curation and quality assurance are needed to create a stable open research data environment. A way of incentivising these activities should be sought.

  Pilot 5 showed that providing guidelines with few, clearly stated and practical suggestions greatly helped different projects to better tailor their dissemination and communication strategy. Focusing more on shaping their target group can also make their strategy more effective. This can be done with different media and was especially tested with specific tweets, as suggested in the OpenUP guidelines.

- **Encourage Open Online Research in research projects**
  Pilot 4 showed that software services developed in order to encourage citizens to gather and analyse data online can give valuable insights from a much wider audience at a small cost. This novel approach to citizen science is feasible under certain conditions: the software needs to be designed and presented in an inclusive way and the technical and software flow needs to be complemented by human moderation to answer questions and resolve dispute. The opportunities for citizen science in the humanities and social science are significant and could be encouraged as much as possible also in future call descriptions.

1.5.2. Developing scholarly publication models

- **Lessons learned from academic piracy for shaping business models in scholarly publishing**
  A similar development like in the music industries since the late 1990's can be observed in scholarly communication. The available technologies enable scholars and other users to rather easily circumvent paywalled content. The advanced Open Access movement led to legal ways to do this, e.g. through gold
or green Open Access. However, as our results showed, many researchers still often use piracy content. We suspect that the convenient one-click access that shadow libraries provide to full digital copies plays a role.

Open access publishing mandates for publicly funded research can create a lucrative revenue stream for publishers in the form of article processing fees. Nevertheless, considering current developments, academic publishers should ask themselves if the copyright and exclusivity-based business models are sustainable on the long run since there is a growing quantity of online tools and services, which do not require paywalled content. In fact, these tools thrive in environments where there are no artificial technical or legal boundaries in front of the accessibility, circulation, consumption of content.

2. OpenUP Pilots Goals & Evaluation Approach

The aim of the pilot studies was to test and/or evaluate selected innovative peer review, dissemination, and impact measuring approaches applied to specific research areas and communities. To achieve this, the OpenUP team has involved and committed interested and engaged research communities from arts & humanities, social sciences, energy, and life sciences. Together with the communities, the OpenUP team applied and tested technical and processual solutions identified in the framework of research of WP3-WP5. Specific stakeholder requirements as well as specificities of the particular peer review, dissemination, and impact settings were addressed as well.

<table>
<thead>
<tr>
<th>OpenUP Pilot Study</th>
<th>Involved Community</th>
<th>Numbers of organisations and individuals involved</th>
<th>Estimated Impact</th>
</tr>
</thead>
</table>
| **Pilot 1: Open Peer Review for Conferences** | European Machine Vision Forum (EMVA)                     | approx. 30 organisations and 100 individuals      | Short-term:  
|                                              | eHealth                                                  | approx. 10 organisations and 30 individuals       | • Awareness and discussions on OPR                                         |
|                                              |                                                          |                                                  | Long-term:                                                               |
|                                              |                                                          |                                                  | • CMS framework for OPR available to the public                           |
| **Pilot 2: Open Peer Review for Research Data** | Human Mortality Database (HMD)                           | 3 research institutions involved                  | Short-term:  
|                                              |                                                          | Interviews with 7 research managers more than 1500 questionnaires completed by HMD users | • contributing to the understanding of users’ requirements through the ad hoc collaborative user survey |
|                                              |                                                          |                                                  | Long-term:                                                               |
|                                              |                                                          |                                                  | • application of transparency good practice to develop data repositories in Social Sciences |
| **Pilot 3: A data journal for the Arts and Humanities** | DARIAH                                                   | 2 contact persons in DARIAN-EU and DARIAH-DE      | Short-term:  
|                                              |                                                          | 20 workshop participants (2 workshops)             | • contributing to the ongoing discussion on data sharing in the Humanities |
|                                              |                                                          |                                                  | Long-term:                                                               |
|                                              |                                                          |                                                  | • implementation of results in the development of a data journal using the infrastructure within DARIAH |
| **Pilot 4: Transferring the research lifecycle to the web (Open Science Repositories)** | Members of Zooniverse Students Developers               | 3 organizations 20+ individuals so far Estimated for live test: 100+ individuals | Short-term:  
|                                              |                                                          |                                                  | • User experience                                                          |
|                                              |                                                          |                                                  | Long-term:                                                               |
|                                              |                                                          |                                                  | • Empowerment of citizens: access to an easy to use collaboration tool     |
| **Pilot 5: Addressing & reaching businesses** | SmarterTogether project (from Zooniverse)               | 3 organisations 6 individuals                     | Short-term:  
|                                              |                                                          |                                                  | • Guidelines applied by                                                    |
This document gathers all final evaluation reports from the individual pilots. In the first chapter, we included a synthesis of the key results and lessons learned from the pilots. Where possible and appropriate, we provided a gender-sensitive evaluation analysing experiences and challenges related to gender and diversity in context of the pilots. In all pilots we attempted to address gender issues, gender balance, and to analyse data by gender. For research activities related to the pilots and involving communities, in particular interviews and user testing, we tried to include different people (irrespective of their gender, ethnicity, career stage).

The insights gained from the evaluation of the individual pilots delivered further input on working practices, developing standards, and remaining gaps. Key insights have been/will be incorporated into the WP3 and WP4 framework study updates (D3.4, D4.3). The pilot studies also provided lessons learned and good/best practices that have been added to the policy recommendations produced in WP7 (D7.3, D7.4). Our synthesis summarises remaining gaps and key insights that we collected together with the involved communities during the test runs of the applied or analysed innovative peer review, dissemination, and impact measurement approaches.

3. Final Evaluation Reports

3.1. Pilot 1: Open Peer Review for Conferences

3.1.1. Introduction

An important aspect of the process of doing science is peer reviewing. It is one of the most important tools for quality assurance of the scientific method. The traditional approach to peer review is single- or double blinded reviewing done by a small group of appointed reviewers. Despite its relevance, traditional peer reviewing has three major drawbacks:

1. It does not scale well. The number of submitted papers at both journals and conferences is steadily rising while the number of expert reviewers is not increasing proportionally.
2. Double-blindness is often compromised as reviewers have a very good overview of the participating research groups and can often infer authorship by language and topics.
3. Compromised double-blindness or single-blindness can introduce a lot of bias in the evaluation of research. Additionally, unfair evaluation can lead to an enforcement of predominant ideas and hinder progress.

Open peer reviewing (OPR) is an alternative, completely transparent approach, which can help solving these challenges. There are multiple ways to open up the review process:

- Open Identity: Authors and reviewers are aware of each other's identity
• Open Participation: A larger community is involved in the reviews
• Open pre-review: early versions of material are public before the review
• Open Report: Review report is published alongside the publication
• Open final-version comments: commenting online possible after the verdict

In Pilot 1, we tested the practicability and impact of OPR at conferences. The first venue for Pilot 1 was the Second European Machine Vision Forum 2017 (EMVA 2017) and the second venue was the eHealth2018 Master Student Competition.

3.1.2. Venue 1: EMVA 2017

The EMVA forum is a conference where researchers from European industry and research institutes come together to present the newest achievements in computer vision and machine learning. It focuses on connecting industry and academia. The open call for papers allows everyone to submit a short paper (2-3 pages) about their topic. Traditionally the conference organizers selected a subset of these submissions to be presented as oral presentations during the conference. The remaining submissions can participate as posters. The conference organisers agreed to test the four OPR principles “Open Identity”, “Open Participation”, “Open Report” and “Open final-version comments” for all papers submitted through the submission system.

The most important changes to the traditional workflow of paper submission, reviewing and voting include:

• All participants can see all submissions after the submission deadline and can discuss the submission with the authors
• Instead of strict reviews, only shorter comments of about a paragraph are used to summarize the individual opinion and suggestions for improvements
• All identities connected to comments and submissions are visible to all participants
• For the final voting all participants have four votes; project committee members get 10 votes.
• The final result based on the sum of all votes from all members was accepted as a final ruling
• All discussions can continue in the CMS interface after the conference has finished

The software to manage the submission, reviewing and voting of potential contributions is called a conference management software (CMS). We had to create our own CMS solution\(^2\) that supports the specific mix of new OPR features needed for Pilot 1 (see “Results”).

In total 78 users registered to the CMS: all persons who submitted a paper through the system (13), the program committee members (8), the PC (1), and additional 56 people who registered after the voting was already over. However, only the submitting authors, the program committee, and program chair (22 in total) used the CMS during the peer review and voting phase. In general, the stakes were not high for this OPR test phase: In total, there were 13 submissions of which ten were chosen to be presentations. The remaining three still participated as posters.

3.1.3. Venue 2: eHealth2018 Master Student Competition

The eHealth conference is an international annual conference for eHealth research and applications. Each year they also hold a student competition which allows master students to participate and submit research papers. Traditionally, a jury used reviews created by assigned reviewers in a single-blind fashion (reviewers could see participants’ names) to select a single winner. For the student competition, we again tested three OPR principles “Open Identity”, “Open Participation”, and “Open Report” but with a different setup:

• After the submission deadline, all submissions stayed hidden. However, each participant of the challenge had to write two ‘lay-man’s reviews’ (“Open Participation”). These initially double-
blind reviews were augmented by traditional ‘expert reviews’ done by external assigned reviewers.

- The rebuttal phase allowed each participant to withdraw his submission based on all the reviews. In this case his contribution would have remained hidden and all involved persons (reviewers and authors) will stay anonymous. All participants moving forward and staying in the race at the end of the rebuttal phase will move to an “Open Identity” status: all submissions, the reviews (“Open Report”), reviewer’s names and author’s names are visible to all conference visitors.
- The program committee used all available reviews (lay-man and expert) to decide on the final winner of the competition.

Again, we had to adjust our CMS software to accommodate this setup. A total of six students entered the competition and eight external reviewers created a total of twelve expert reviews. No contestant opted out of the competition and all submissions entered the “Open Identity” phase later on. The program committee took all available reviews (two lay-man and two expert reviews for each entry) into consideration. The final winner was announced during the last day of the conference.

3.1.4. Results

3.1.4.1. Open Source Conference Management Software Repository

The features needed to conduct the proposed OPR process were not available in existing conference management software at the project start. Thus, a dedicated CMS version with OPR workflow support was created based on the popular HotCRP CMS code base. The resulting source code has been released to the public under an open source license at the repository https://github.com/mthz/hotcrp.

3.1.4.2. EMVA Feedback Survey

The results on the feedback from 19 of the 22 participants showed a very positive response. Overall, the participants expressed a strong acceptance of the proposed OPR process.

- Clearness of execution and implementation/CMS interface: 94% strongly agree
- Acceptance of OPR (review phase): 100% strongly agree
- Acceptance of OPR (voting phase): 88% strongly agree
- Open identity skewed feedback towards too much positivity: 69% agreed
- Overall acceptance of OPR approach, would support again: 94% strongly agree
- Greatest fear associated with OPR included: biased/whitewashed reviews due to non-anonymity; backlash for bad reviewing (maybe over other channels/private email etc.)

In addition, the team gathered personal feedback from the CMS users. Many indicated that they liked the reduced effort associated with the commenting system and in general appreciated the inclusion of all authors into the voting process.

3.1.4.3. eHealth2018 Student Competition Feedback Survey

The feedback survey was sent to all students, external reviewers and the program chair who were involved in the peer review process. Six of the 15 people using the CMS returned the feedback survey. The overall response was positive with some reservations by the authors regarding the added effort of having to do extra work (lay-man reviews).

- Clearness of execution and implementation/CMS interface: 83% agree
- Acceptance of OPR (lay-man reviews): 50% strongly agree
- Acceptance of OPR (open identity): 67% strongly agree
- Open identity skewed feedback towards too much positivity: 17% agreed
- Overall acceptance of OPR approach, would support again: 83% strongly agree
- Greatest fear associated with OPR included: Lay-man review forces people to review things outside their expertise (this makes the reviewer look bad).
Again, we collected qualitative feedback at the conference itself and had discussions with many of the involved participants of the competition. During these talks, the lay-man reviews remained a controversial topic: some felt that it offers too little value for the time invested while others feared, that the combination of enforcing everyone to review while revealing identities at the end, might lead to retribution and the possibility of creating a bad reputation within the community. On the other hand, having the possibility for doing reviews without too much pressure was perceived by most students as a valuable way to get experience in the scientific method.

3.1.4.4. Presentations and discussions at conferences

Posters and extended abstracts outlining pilot 1’s setup, goals, and early results have been presented at multiple major open science conferences:

- Open Access Tage 2016 (2016-10-11, Munich, Germany)
- Munin Conference on Scholarly Publishing (2016-11-21, Tromsø, Norway)
- OpenSym 2017 (2017-08-23, Galway, Ireland)
- Munin Conference on Scholarly Publishing (planned: 2018-11-28, Tromsø, Norway)

Presentations and discussions with conference attendees were used to create more awareness for open peer review and the possibilities of using OPR for conferences. Furthermore, inputs and suggestions by the communities were included in the setup process of our pilot.

3.1.5. Lessons learned

Feedback from the researchers involved in the OPR process at the EMVA conference and eHealth 2018 student competition was positive. Overall, the participants expressed a strong acceptance of the proposed OPR process and would support it again. The participants' greatest fears associated with OPR included: biased/whitewashed reviews due to non-anonymity; backlash for bad reviewing (e.g. over other channels/private email); and added effort and risk for reviews outside one’s own expertise (layman reviews). Also, the conference organisers of the EMVA are willing to continue applying the OPR approach for the next conference.

3.2. Pilot 2: Open Peer Review for Research Data in Social Sciences

3.2.1. Introduction

Similarly to peer review of publications, data peer review is a quality assessment process of a dataset performed by experts in the field. Data quality assessment is a complex process that has to consider the different phases of the data lifecycle, starting from the development of a data management plan (DMP) at the initial stage of a scientific project to the publication of its results. Data publication (some authors speak about Publication with capital letter, Lawrance et al. 2011, Mayernik et al. 2015) that undergoes a peer review process validating data quality is currently performed in data journals and data repositories. However, analyses by Candela et al. (2015), Carpenter (2017) show that there is room for improvement, as peer review activities in data journals vary widely and are mostly focused on metadata rather than data themselves, aiming at assessing the documentation and metadata description that facilitates data reuse. Assante et al. (2016), in the analysis of generalist repositories (Zenodo, Dryad, Figshare etc.), also come to the conclusion that different criteria and quality control mechanisms are implemented based on various policies and/or guidelines. Best practices of data validation can be taken from the publication in trusted data repositories (Callaghan et al. 2014), in which the data quality control can be considered a form of review by experts in the field carried out in a pre-publication phase. Moreover, the possibility to post comments and evaluations by end users may be considered as a trait of open participation in Open Peer Review (OPR) performed in a post-publication phase. Other
important indicators of impacts can be developed using data citation counts and/or statics of use, which are considered a crucial aspect to improve data sharing. Pilot 2 was devoted to analysing OPR for research data in the scientific community of the Human Mortality Database (HMD). Table 1 provides a brief description of HMD.

The detailed methodology adopted (see for reference OpenUp D6.1 and D6.2) comprises two parallel activities. The first one is the set of interviews to HMD managers and country specialists, who are responsible for the validation of data coming from the contributing countries. These interviews provide us with important information on procedures performed to assess the quality of data in a pre-publishing phase as well as strong and weak points connected with data sharing. The second activity, carried out in collaboration with HMD management, is the development and the submission of a questionnaire to HMD users. The results of the survey help understanding the users’ practices in data access and use that can be considered as proxy indicator of post-publishing appreciation of the quality of the database.

Table 1. The Human Mortality Database in a nutshell

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Human Mortality Database (HMD) is an open database that provides detailed, consistent and high quality data to researchers, students, journalists, policy analysts, and others interested in the history of human longevity and its prospects for the future (<a href="https://www.mortality.org/">https://www.mortality.org/</a>).</td>
</tr>
<tr>
<td>HMD is a joint project of the Department of Demography at the University of California, Berkeley (UCB), and the Max Planck Institute for Demographic Research (MPIDR). Recently the HMD project benefits also from a continuous support by the French Institute for Demographic Studies (INED).</td>
</tr>
<tr>
<td>Along with raw data, coming mostly from national statistical offices, HMD provides uniform death rates and complete and abridged period life tables. In addition, cohort life tables are provided when the observation period is sufficiently long to include at least one cohort observed from birth until extinction. All data are provided with the highest level of details and include some unique information on old age mortality up to age 110.</td>
</tr>
<tr>
<td><strong>Documentation available:</strong> Methods Protocol - Country specific documentation - Guidelines for citation - User agreement - Citation report</td>
</tr>
<tr>
<td>Country specific documentation describes in depth all necessary information to understand the population dynamics as well as the issues related to the computation of the raw data. It also discusses any data quality issues that might arrive from the original statistics. This report is updated each time new data are analysed.</td>
</tr>
<tr>
<td><strong>Type of data:</strong> Original input data - Period and cohort life tables - Unsmoothed death rates - Population estimates - Death counts - Life expectancy at birth and all other ages.</td>
</tr>
</tbody>
</table>

3.2.2. Interviews

The interviews were conducted on the 31th of January and 1st of February 2018 at the Max Planck Institute for Demographic Research in Rostock, Germany. They were performed according to interviewees’ role in HMD, and for the selection of interviewees we considered gender balance. As researchers in this area are prevalently male, women are also underrepresented in this sample. The two directors, two researchers in their role of country responsible (in charge of analysing data for specific countries) were interviewed.

These interviews covered the majority of HMD staff (4 out of 7). Moreover, in Rostock we also interviewed a HMD user to test the questionnaire to be further submitted online to end-users. Interviews to HMD staff were planned aiming at exploring the following topics:

- *Origin, motivations and organisational features*

---

4 Off records and in informal discussion with the HMD female director we touched upon gender aspects. She mentioned her experiences with having responsibilities and collaborating within a prevalently male scientific community. She is generally one of the few women in meetings and conferences, but she did not register strong discrimination. Competences, capacity and passion have not been stopped by gender prejudices in her case.
In the following a brief summary of the interviews is provided considering the main issues that emerged.

### 3.2.3. Origin, motivations and organisational features

HMD was launched in 2002 as a result of a collaborative project that involved researchers of the Department of Demography at the University of California Berkeley (UCB) and the Max Planck Institute for Demographic Research (MPIDR). Two previous relevant experiences guided the development of the database: the Kannisto-Thatcher Database on Old Age Mortality (KTD) at the MPIDR and the Berkeley Mortality Database (BMD), founded by John Wilmoth at UCB. Both experiences were concerned with what was at that time an emerging phenomenon of low mortality at young and adult ages, falling mortality at old ages, and greater survival to an advanced age, leading to a potential increase in the number of people exposed to degenerative diseases, which are difficult to treat or prevent. To understand this phenomenon, it was necessary to analyse and model longevity and survival of humans with a special emphasis on advanced (frontier) age over a long period of time. This research needed reliable data at international level providing long-term and continuous series without gaps, running up to the highest ages, providing fine details according to age, time, and cohort dimensions, ensuring sufficient quality and comparability across time and populations. HMD was therefore developed to answer this scientific question providing a methodology based on the previous mentioned experiences as well as freely available high-quality data.

The collaboration was originally based on a small, very well-established group of internationally based demographers who were willing to serve the scientific community interested in demographic studies. Work is equally distributed among the team that comprises country specialists (CSs), who have high level competences on demographic development of a set of specific countries and are responsible for collecting and analysing data from the related national statistical offices. Other tasks comprise the development of computer codes, which are also made freely available to the end user who wants to reproduce the analysis, as well as the management of the website. Strong collaboration pertains the data quality process performed before data are publicly available, which constitutes a form of internal pre-publishing peer review process. Trust among the team and scientific curiosity are the drivers of this successful cooperation, that only recently was formalized by a Memorandum of understanding.

### 3.2.4. Goals and main features of the database

The main goal of HMD is to support research on human mortality and longevity providing open data on 39 countries and some sub-areas and sub-populations with series starting as early as 1751 (i.e. Sweden) and covering more than 100 years for 16 populations.

Birth and death counts are generally based on data from national vital registration systems, while data on population are based on the national census and estimates between censuses. However, differences may exist among countries in the periodicity of census, methods and definition used as well as in data format. Moreover, some countries have experienced changes in their territorial boundaries, have suffered substantial loss during war periods and or faced substantial consistent migration over the period covered by HMD. For these reasons HMD has developed a methodology to produce detailed death counts and population estimates, to correct mortality estimates at old ages, and to build high quality life tables (as described in detail in the Methods protocol). All HMD data are prepared using this standard methodology. This assures comparability in time and across countries. When special methods are needed to accommodate issues in data availability, this is documented in the country-specific documentation as well as reported in summary table5. Country-specific details related to the data quality

5 [https://www.mortality.org/Public/Docs/SpecialMethods.pdf](https://www.mortality.org/Public/Docs/SpecialMethods.pdf)
and statistical system in each country are therefore documented in the country-specific Background and Documentation file accessible from each country webpage. The application of these thorough procedures, the punctual explanation of the estimations and refinements of data sources make this database different from other sources providing mortality rates. These procedures guarantee a uniform analysis of raw data, facilitating the comparability across time and space, while the detailed documentation and the availability of source data allow end user to reproduce the analysis.

The HMD team has also developed software code that guide them in the evaluation of data quality as well as software packages that facilitate end user to import and working with HMD data. These tools are freely available to end users along with technical reports explaining how to use these scripts\(^6\). This is another value-added feature of HMD.

3.2.5. Data quality assessment process

The HMD team has developed a set of procedural steps to ensure data quality. This important topic was addressed in the interviews with the two directors and particularly explored in the interviews with the country specialists (CSs). An activity diagram that reconstructed the workflow of the activities performed before data publication was presented to the CSs and discussed to have further insights on the procedures adopted to assess data quality. This intended to explore whether collaborative activities resembling a peer review process could be tracked in HMD data quality assessment.

A high-level description resulting from the interviews is provided in Figure 1.

**Figure 1. Data quality assessment process**

Each country or area is assigned to an individual researcher, a CS, who maintains a close relationship with a local expert generally at national statistical offices, has an extensive knowledge of the population dynamics as well as how data are collected at national level. A CS is responsible for the first quality checks that evaluate consistency and plausibility of input data, prepares pre-calculation file (Lexis files)

\(^6\)https://www.demogr.mpg.de/en/projects_publications/publications_1904/mpidr_technical_reports/all.htm
and analyses the results on the basis of a pre-defined data quality checklist and diagnostic charts that help him/her to explore unusual fluctuation and/or any other issues in data sources. The results of this analysis are shared within the HMD community via an internal report and are the basis for the application of the six-step procedure to produce the complete data series (exposures to risk, death rates, life expectancy and other life tables). Before data are published, the HMD team perform an additional phase of validation. These activities are crucial especially when a new country has to be included in HMD. However, they constitute a routine procedure every time data are updated. In cases of unexpected changes in national statistical systems or in regimes of national statistical registration, the updating procedures are non-trivial.

All steps in the computing of data analysis are documented in detail and made available to end users in the different files (Background and documentation, Data source and Explanatory Notes). According to the CSs interviewed, this is the distinctive feature of HMD. Data refinements and harmonization that allows comparison across countries are documented in detail so that researchers in this field are aware of possible problems in the data and know how these issues have been solved.

3.2.6. Opinion on open access of data and peer review

The HMD management team declared that open access and open data in particular are very important for the development of demographic studies. Although they have no official statements on open policy, since its beginning, HMD provided open access data, based on a user agreement indicating that “the data in the Human Mortality Database (HMD) are provided free of charge to all individuals who request access to the database”. Moreover, users are required to cite the database in their publications, following the citation guidelines provided by HMD. Citations tracked through Google scholar are also reported in the website, and further steps to improve their collection are going to be planned in the next future.

When asked about long preservation of data, it emerged that they are dependent of funds. At the moment MPIDR support their activities (MPIDR researchers are allowed to spend half of their work time on HMD), while the UCB team has to provide its own funds. A clear commitment of the organisation would therefore be very important and would also mean a clear recognition of their activities. Between the lines, it emerged that publication of scientific papers are generally considered more important than managing a database. In their opinion, the analysis of data, their quality check is not only a service for the community of reference but is a researcher activity in itself.

The majority of the interviewees has heard about open review of journals but has little knowledge on all its traits. If they see a similarity with open peer review of data, this is associated in particular with transparency as a means of reconstructing the methods and procedures used for the data analysis.

3.2.7. Survey of the HDM users

The survey was based on an online questionnaire using LimeSurvey, an open source software that also supports invitations, reminders, and makes answers anonymous. The questionnaire was developed in collaboration with the HMD managers through web calls and a face-to-face meeting in Rostock. A pilot version was created and disseminated to expert colleagues. Comments from experts were then incorporated into the final version of the survey. Finally, questionnaires were sent to a small number of respondents (n=10) to pre-test it. The survey period was March - June 2018.

The survey made use of a semi-structured questionnaire of 20 questions, most of them were multiple choices, while plain text answers were also included to collect researchers’ opinion on specific features of HMD database.

The questionnaire consisted of two main parts. In the first one, respondents were asked for information on sex, age, country of residence, occupational position and related institutional affiliation as well as

---

<https://www.mortality.org/Public/CitationGuidelines.php>
main field of interest of their work. These elements allowed us to provide a more detailed profile of the HMD users.

In the second part, questions were specifically focused on the HMD user's practices and attitudes in data access and use. In particular we explored the following aspects:

- General information on access (frequency and length of use - countries of interest)
- Modes of dataset acquisition (manual or automatic downloads - type of datasets)
- Dataset use (purpose in using - ways of processing dataset - software used - other source of information used in the field)
- User's perception of HMD (advantages - comments and suggestions)

Data in HMD are provided free of charge, but users are required to register and accept an HMD user agreement. Therefore, questionnaires were sent to all individuals who have registered to HMD website\(^8\). This sample is able to represent the whole target group who uses HMD data, even if the number of registered users may be biased by multiple accounts and changes in mail address.

35512 invitations have been sent, 1049 came back for incorrect address, 1553 completed the questionnaire. The response rate was 4.5%.

3.2.8. Results

3.2.8.1. Respondent's profile

An overview of the respondent's profile is given in Table 2. The majority of respondents are male (68.8%) and they fall mainly into two age groups (from 20-39=43.7% and from 40-59=38.6%). Most female respondents (29.5%) fall in the same age groups (from 20-39=16.9% and from 40-59=10.5%). Respondents residing in Europe comprised 59.9% of total responses, followed by America (25.3%), Asia (10.6%), Oceania (2.9%), and Africa (1.2%).

60.2% of respondents work at research institutions, distributed by University (51.0%), Other public training or research organization (7.1%) and Other private training or research organization (2.1%). 13.4% are employed in Insurance/Re-insurance companies. Of the female respondents, most work at Universities (16.0%), at Insurance/Re-insurance companies (3.5%), and at Other government organization, statistics office (3.1%). These data are confirmed by the responses about the occupation, in which the majority of users are researchers/scientists (32.2%, of which 9.9% are female), teachers/professors (20.1%, of which 4.8% are female), students (14.6%, of which 6.6% are female) and actuaries (19.1%, of which 4.8% are female).

As expected, the main fields of interest are demography (25.1%), actuarial studies (29.6%) and statistics (11.5%).

Table 2. Respondents' profile

<table>
<thead>
<tr>
<th>Respondents' profile</th>
<th>No</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1069</td>
<td>68.8</td>
</tr>
<tr>
<td>Female</td>
<td>458</td>
<td>29.5</td>
</tr>
<tr>
<td>Prefer not to answer</td>
<td>26</td>
<td>1.7</td>
</tr>
</tbody>
</table>

\(^8\)The survey questionnaires were sent by the HMD team from the HMD server at UC Berkeley. The OpenUP team had no access to the personal data of HMD users and were simply sent the anonymous answers to the survey.
If we consider how respondents have learnt about HMD database, a consistent number of the answers (29%) indicate that it was mentioned by a colleague or found in a web search (21.1%) or cited in an article (16.7%).

### 3.2.8.2. General information on access

Considering length of HMD use (Q.9), the majority of responses (55.9%) registered less than 5 years and less than a year, while 34.9 are long-standing users (less than 10 years and 10 years or more). 9.3% of respondents declare that they never used HMD after registration, therefore they have not completed the remaining questions. Thus, the analysis of further questions is based on the sample of current users, that is 1408.

<table>
<thead>
<tr>
<th>Age</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;20</td>
<td>4</td>
<td>0.3</td>
</tr>
<tr>
<td>20-39</td>
<td>679</td>
<td>43.7</td>
</tr>
<tr>
<td>40-59</td>
<td>600</td>
<td>38.6</td>
</tr>
<tr>
<td>60+</td>
<td>270</td>
<td>17.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country of residence /continent</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>19</td>
</tr>
<tr>
<td>America</td>
<td>393</td>
</tr>
<tr>
<td>Asia</td>
<td>165</td>
</tr>
<tr>
<td>Europe</td>
<td>931</td>
</tr>
<tr>
<td>Oceania</td>
<td>45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of Institution</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>University</td>
<td>792</td>
</tr>
<tr>
<td>Other public training or research organization</td>
<td>110</td>
</tr>
<tr>
<td>Other private training or research organization</td>
<td>34</td>
</tr>
<tr>
<td>Other government organization, statistics office</td>
<td>136</td>
</tr>
<tr>
<td>International organization (United Nations, World Bank, etc.)</td>
<td>19</td>
</tr>
<tr>
<td>Insurance/Re-insurance company</td>
<td>208</td>
</tr>
<tr>
<td>Other large private corporation</td>
<td>44</td>
</tr>
<tr>
<td>Other Small and Medium-size private organization</td>
<td>92</td>
</tr>
<tr>
<td>Foundation</td>
<td>6</td>
</tr>
<tr>
<td>Other non-profit/NGO</td>
<td>29</td>
</tr>
<tr>
<td>Other</td>
<td>83</td>
</tr>
</tbody>
</table>
Related to the frequency of access the database (Q.10), the majority of respondents (76.1%) consulted it a few times over the past (45.2%) and rarely (30.9%). This can depend on the types of data and but also by their updating that it is generally done at 2- to 3-year intervals. (Barbieri 2015).

When asked about country of interest (Figure 2), an high number of respondents (45.6%) report that they access data related to all countries available in HMD, while the most accessed data are the ones of the U.S.A (22.2%), followed by the U.K (17.8%), Germany (15.4%), France (13.3%), and Sweden (10.9%).

**Figure 2. Distribution of respondents by question: “Which HMD countries/regions are you most interested in?”**

3.2.8.3. Modes of dataset acquisition

The second section investigated the types and acquisition mode of HMD dataset (Q.12, Q.16).

A specific question (Q.12) about data acquisition mode leads to the distinction between registered users who usually only consult HMD data and those who download and/or copy files from HMD website. This is a first indicator of data usage. 12.1% of respondents affirm that they never download/copy files, the rest declare that they only select data from HMD website (85.5%), 7.1% automatically download data using some computer codes, while the remaining 7.4 % use both modalities.

To explore which types of dataset are more used, respondents were asked about their preferences (multiple choices were allowed). Figure 3 shows that 65.1% of all respondents indicate life table, followed by death counts (40.9%), life expectancy at birth (36%), population estimates (34.2%), unadjusted death rates (27%), cohort data (23.5%) and zip file of pooled HMD data (18.3%). It is interesting to note that only 7.3% of respondents usually access input files. As mentioned above, these are the baseline data on which HMD results are computed. This is a probable indicator of the reliability of HMD data, as users usually do not have the need to access input data to reproduce the analysis.
3.2.8.4. Use of dataset

The third section of the questionnaire analyses why and how HMD data are used (Q.14, Q.15, Q.18 and Q.19).

Considering the purposes of accessing dataset (multiple choice allowed), 36.7% use HMD for research in mortality, 23.1% for educational purposes, 22.8% to also monitor mortality trends, while 11.6% indicate that they use data for business activity. This is likely to reflect the composition of HMD user profiles.

HMD database shared data contents in ASCII text files, imported into Excel tables, or into a statistical package (e.g., R, SAS, Stata, SPSS, etc). Figure 4 shows that when respondents were asked on the type of software used to process HMD data (multiple choices were allowed), the most frequent answers are Excel (62.3%) and R (48.9%). Also in this case, the usage of statistical packages may be considered an indicator of expert users.

Additional questions were focused on respondent’s practices in using HMD data. The first one asked whether and how they elaborated HMD data to conduct their further analysis (Figure 5). Multiple answers were allowed. As most frequently reported, HMD is the basis for statistical modelling (43.7%), demographic forecasts (37.8%) or for the identification of additional indicators (36.7%). Some respondents also combine HMD data with other sources (42.1%).
When asked which other information sources are consulted, respondents report that they also access on regular basis data provided by National Statistical Offices (44.5%) as well as by International organizations such as the WHO Mortality Database (30.5%), the United Nation Population Division (25.7%), the Eurostat (25.6%), the Centers for Disease Control (including the NCHS) (18.2%), the World Bank (17.4%), the US Census Bureau International Database (11.4%) and the Institute for Health Metric and Evaluation (4.4%).

3.2.8.5. User’s perception of HMD

The last section of the questionnaire (Q.17 and Q.20) intends to explore users’ perception on the advantage in using HMD data. Respondents reported most often the following options (Figure 6): the easily accessible data (59.4%), the comparability over time and across-countries (55.5%), and the long-time periods of the data available (45.2%).

Moreover, looking at the final overall comments provided by some respondents, important indications on users’ needs and expectations can be drawn. Among the suggestions of improvements, there is the request of providing a more detailed geographic distribution of data (e.g. provinces, municipalities),
expanding the number of countries to be included in HMD, including data on causes of death. Different needs emerged on the file formats, depending on type of users and purposes. The most frequent demand concerns the improvement of tools for data automatic imports into statistical packages, such R and Stata. Even if some tools have been developed by the HMD team, it is clear that some users are not aware of them. Therefore, there is room for improvements in the HMD web interface, making the link with these scripts more evident. Moreover, respondents mention the need of a more frequent timely update of dataset.

Among the many appreciations reported by respondents, some of them summarize well the characteristics of HMD. These comments could be also relevant to increase data publication and quality assessment in other Social Sciences contexts. This pertains the availability of the source data, the transparency of in the data processing procedures, the detailed documentation and the trustworthiness of the database. As a respondent reported "It is also easy to use and reference, and a trustworthy source, so I don't have the need to look elsewhere for data". Another respondent expressed his/her appreciation: "You are the gold standard in the field and an example of the good work that can be done, but we need more like you to have the rest of the world a la HMD".

3.2.9. Lessons learned

Some important indications emerged from the analysis of the interviews that can drive the adoption of data quality assessment, and hence peer review, as well as some principles that can incentivize other scientific communities to share their research data. As stated by the HMD interviewees, the guiding principles to create an open access database were: comparability, flexibility, accessibility and reproducibility. Comparability was reached using a uniform, scientific methodology to calculate the various statistics of the 39 countries included in the database. Flexibility was achieved in the analysis of results using a uniform set of procedures for each population, but at the same time giving significant attention to each population in terms of its history and socio-political development. This is also reflected in the available formats of output data series. This is achieved thanks to the experiences and knowledge of country specialists, that is persons in charge of collecting data from a specific number of countries, who interact with statistical offices, check data consistency and provide population statistics together with a country report that explains specificity and motivation of analysis. Accessibility was guaranteed from the beginning by free of charge access of data, as well as by the provision of data in an open, non-proprietary format. Reproducibility is provided by the reconstruction of the data lifecycle that includes the availability of raw data, the method applied, the related results as well as the explanatory documentation. One of the main successful features of HMD is its transparent way of data managing and sharing that has two central phases of data validation. The first one is carried out by the CSs, who analyse the raw data according to a common predefined checklist that verifies consisten and plausibility of data. The second one is carried out in a collaborative way within the HMD team that validate the statistics before their publication, each time the database is updated.

Moreover, another successful component of HMD was its collaborative approach that is based on a strong scientific interest in the field as well as on the trust among the involved community that only recently has formally signed a Memorandum of understanding.

The interviews also highlighted some indications that confirm some concerns already mentioned by other surveys. Interviewees stressed the importance of having a strong commitment of the organization in supporting the development of data infrastructures. This pertains different aspects: a long-term financial support (beyond the project duration), a policy endorsement on open data as well as a formal recognition of scientists for the efforts in data curation and quality assurance.

Considering the results of the survey, users confirm the main strength points of HMD regarding in particular the accurate and well-documented data quality assessment that make the process transparent and facilitate the reproducibility of the analysis. They do not outline evident weak points;
they rather suggest improvements mainly related to the provision of tools that facilitate the import of data into statistical packages. This may be also related to a simple style interface, where some links could be better highlighted. A user’s comment summarises well this aspect: “the format of the website could be more aesthetically appealing, but as it is the site is very functional and suits the needs of the users”. Moreover, the different types of user profiles that comprise the research field as well as the private sector, addressing different users’ needs are indications of the importance of data sharing that reinforce Open Science principles. If considered under an OPR perspective, a straightforward transposition of the procedures adopted for scientific journals seems to be hard to apply. However, some traits of OPR, such as transparency in the quality assessment process, represent for open data a feature that should be promoted at a larger scale. This could be also applied to the trait of open participation that in case of open data implies a more common use of data citations by end-users as well as the implementation of additional tools to track data re-use. Further research is needed to explore practices of data sharing and management not only in Social sciences, to take the necessary steps to support and improve high quality data sharing.

3.2.10. References


Lawrence Bryan, Catherine Jones, Brian Matthews, Sam Pepler, and Sarah Callaghan (2011). “Citation and peer review of data: Moving towards formal data publication.” International Journal of Digital Curation, 6(2): 4–37. doi: http://dx.doi.org/10.2218/ijdc.v6i2.205


3.3. Pilot 3: A data journal for the Arts and Humanities

3.3.1. Introduction: Data publishing in the Humanities

Within the digital ecosystem the flow of information and innovative practices, such as data mining, data harvesting and commenting slowly transforms Humanities research. Research practices in the Humanities increasingly incorporate the use of digital resources and tools. As open access becomes a vital aspect of research dissemination and of the scholarly discourse in general, an increasing proportion of research results in the Humanities are published in an open way. Furthermore, besides publishing
through traditional venues, such as journal and monographs in print or digital form, new forms of publishing are used by researchers (pre-prints, depositing datasets, writing scientific blogs) (Operas, 2018).

Openness permeates every aspect of the scholarly communication process: not only dissemination, but also reviewing and assessing research outputs. The concept of open peer review (OPR; defined and discussed in D3.4 (Görögh, 2018)) is an essential element of Open Science and closely intertwined with alternative forms of publishing, as it is connected to specific features of digital publishing in the Humanities (e.g. annotation technologies and commenting).

The pilot has a dual purpose: on the one hand, it describes the data sharing practices within Humanities research, and on the other hand, the study evaluates how quality assessment and (open) peer review can be applied to research data within this field of study. Based on existing e-Infrastructures and practices of Humanities research groups the pilot analyses and demonstrates the feasibility of a basic workflow that will combine the publication of data with commenting and reviewing systems. The research setting is provided by DARIAH-EU and DARIAH-DE, their extended network of Humanities research groups and by the research groups related to the Campus Labor at the University of Göttingen.

3.3.2. Methodology

The study builds on desk research based on reports and survey executed by DARIAH projects. DARIAH is a pan-European infrastructure for Arts and Humanities scholars working with computational methods. Besides supporting digital research, it puts an emphasis on teaching of digital research methods. DARIAH is also a network connecting several hundreds of scholars and dozens of research facilities, which provides digital tools for researchers and shares data as well as know-how. One of the priorities of DARIAH-EU is to build relationships and engage with communities about research and education based on practice or career level, and to identify user needs within those communities. DARIAH-DE, as an active participant of the DARIAH-EU network, contributes to the international research competitiveness and connectivity by developing a modern research infrastructure for the humanities and cultural sciences. Thus, the gathered information on user needs (DARIAH survey results), and the infrastructural development within the DARIAH-DE project provide the basis for the framework of the data Humanities data journal.

Other inputs for the study were provided by workshop results: 1. OpenUP workshop on open peer review “Open Peer Review hands on: alternative methods of evaluation in scholarly publishing” at the DARIAH annual event on 23 May 2018 in Paris, France, and 2. FOSTER/OPENUP joined training day on open peer review on 20 June 2018 in Göttingen, Germany.

As Digital Humanities (DH) has become an area of Humanities research where innovative tools, digital workflows and new methodologies based on collaborative and interdisciplinary work are employed, more attention is drawn in this field to the development and implementation of new principles and standards that ensure openness, interoperability and processability of scientific information. Therefore, researchers in the Digital Humanities Department at the University of Göttingen and members of the Campus Labor were asked to provide input in connection to data management and data sharing within their research practices.

3.3.3. What is data in the Humanities?

Due to the explosion in the creation and sharing of digital sources in multiple formats, several issues of data management and data exchange should be examined. These issues address the problems all Humanities research groups face in the strengthening European discourse of open data: the common translations of data, common vocabularies for describing data, the transparent process of data management, the legal and infrastructural issues of reuse of data and long-term data preservation. Data, in its more general sense, as something to be measured, collected, reported, and analyzed, is not a common concept in Humanities since most of Humanities content and research cannot be related to numbers and statistics (Konnikova, 2012). However, research data is as diverse as the disciplines that
use them, and they can be defined, on broad terms, as the evidence used to inform or support research conclusions.

However, due to the increasing attention to the research processes and their underlying structures in science, data have become a common basic principle in scholarly discourse including the Humanities fields. It is hard to define research data in the context of Humanities. A definition offered by Schoch is: data in the humanities could be considered a digital, selective, machine-actionable construction of the object of humanistic inquiry. There are two basic types of data in the humanities: big data (relatively unstructured, messy and implicit, relatively large in volume, and varied in form) and smart data (semi-structured or structured, clean and explicit, as well as relatively small in volume and of limited heterogeneity) (Schoch, 2017). Due to the varied fields of study involved in Humanities, the contextualization of data is needed to understand research data management. Obviously, digital humanists would have a different take on what should be considered data and how to store than historians or archaeologists. Therefore, this pilot does not wish to untangle the extremely complex domain of Humanities data management. As the focus of the study centres more on publishing data rather than on data management services, the pilot is more concerned about the ways underlying data can be shared among researchers or published.

3.3.4. What is data publishing?

Publishing data has been defined as a process to make data as permanently available as possible on the Internet (Lawrence, 2011). However, the publication of data is not only making datasets available on the net, but it involves a service of checking the metadata and format of datasets (Callaghan, 2013). Formal data publication usually provides the data user with associated metadata, a persistence identifier for easy discoverability, and a platform for the dataset.

The data journal, defined by the RDA-WDS Publishing Data Workflows Working Group (WG), is a journal (invariably Open Access) that publishes data articles. The data journal usually provides templates for data description and offers researchers guidance on where to deposit and how to describe and present their data. Furthermore, the data article has the primary aim of providing a formal route to data-sharing. It is the carrier of information on the standards of curation, formatting, availability, persistence or peer review of the dataset (Austin, 2017).

3.3.5. Data sharing

Schöpfel and Prost conducted a campus-wide survey on data management, data reuse and data sharing issues in 2015. The anonymous online questionnaire, disseminated within the whole research community on the social sciences and humanities campus (1,800 persons), did not have a high response rate (receiving 270 responses - 15% of the whole sample). The survey provides an insight into the attitudes of students, scholars, research management administrators toward open data and data sharing.

Most respondents (64%) do not share their research data with colleagues or other people. Nobody, except themselves, have access to their data files. Only 34% share data with their colleagues and/or members of their research team, and only 5% share them with a wider audience, in the strict sense of an open data approach. Few replied to the question about the reuse of data produced by other researchers (58%), and even fewer (38%, i.e. 22% of the whole sample) said that they had already downloaded this kind of scientific output. Also, very few answered that they were not interested or motivated to reuse other researchers’ data. One third (38%) simply were (are) not aware of this opportunity and way of doing science.

The general willingness to deposit and share their data was also tested. About 40% of the respondents express a positive opinion about data sharing. Either they have already deposited their data in a

---

9 https://www.sheffield.ac.uk/library/rdm/whatisrdm
repository (16%) or they intend to do so in the future (25%). 30% admit that they were not aware of this possibility.

Another third (29%) dearly say that they never deposited their research data in the past and that they have no intention of doing so in the future. A deposit in a data repository is no option for five reasons: sensitive and confidential data, risk of plagiarism, workload, data illegibility ("nobody would understand my raw data") and intellectual property. This rejection of data sharing significantly decreases when it comes to the question of publishing data along with an article, i.e. when data sharing is incentive or obligatory. 44% of the respondents state that they have already published data together with an article, and 31% announce that they intend to do so in the future while 18% admit that they simply did not know about this possibility.

Regarding the kind of data they would share, only 12% clearly state that they will not deposit or share their research results in this way. 37% underline that they would share those data asked for by their peers while others say that they would deposit data produced in collaborative research projects (33%) or with public funding (25%). As for the infrastructure they would use for depositing data, most of the researchers indicated the use of international data repositories (47%) as a primary preference, followed by local servers hosted by the laboratory, i.e. the research institute (39%). National (35%) or campus-wide platforms (31%) are less preferred. In other words, this sample is more interested in impact and visibility (international repository) and disciplinary specificity (laboratory) than in a multidisciplinary, national or institutional solution.

A web-based survey conducted by the Digital Methods and Practices Observatory (DiMPO), a working group under VCC2 of the DARIAH research infrastructure (Digital Research Infrastructure for the Arts and Humanities) asked European researchers (2177 respondents) about their scholarly practices and digital needs in the arts and humanities. The results show similarity with the Schöpfel survey. A majority of the respondents (9/10) indicated that they use a word processor for storage and management of research assets. Three out of five respondents stated they use spreadsheets, while about one third said they use database management systems or note-taking and bibliographic citation management applications. Only one out of seven presently use web-based content management systems (CMS) to store and manage research assets.

As opposed to their actual practices of sharing data, researchers rated the findability and access to existing digital research resources or data as the most important (scores 9.5/10). A slightly lower score of 9 was granted to digitization of research resources or data currently not in digital form by three out of four respondents. Two other needs, improved findability and access to digital tools or software, and networking with other researchers, research groups and institutions, share third place with a score exceeding 7 by more than three out of four respondents.

3.3.6. Barriers to data sharing in the Humanities

There are numerous barriers researchers encounter in data exchange processes. One of the main obstacles is the generally closed world of scientific discourse in the Humanities. Much of the work in European Humanities research is not visible. The disperse research communities often fail to connect to one another because of the language barriers. Humanities scholars very often publish in their national languages, and the trend is to continue doing so in the future. Europe lacks an integrated database of published journals in various national languages. A database of this kind could be a sort of 'who's who' within a particular field of research.

Another barrier relates to the actual research data. Due to lack of standards and common guidelines in data management, it is very difficult to connect data. There are initiatives on a EU level, which work toward a more unified Humanities research landscape. CLARIN, the Common Language Resources and Technology Infrastructure, is focused on integrating language data across Europe. DARIAH, the Digital Research Infrastructure for the Arts and the Humanities, is more focused on increasing the visibility at the European level of national research related to cultural heritage, digital arts, etc. These two projects

10 https://zenodo.org/record/260101#.W3aPpLpuLQQ
provide a positive direction of development in this field. Both projects try to fill in the gaps where no data exists and try to connect data where it does exist but lives a life of its own in an unconnected place. Since there is a global trend towards inter-, multi-, and transdisciplinarity, the Humanities research field should be encouraged to participate in data exchange processes across research domains.

3.3.7. Current data sharing standards

The basic principle behind developing standards for sharing research data is the purpose of publication of scientific information is to move science forward. In this sense, an author’s obligation is not only to release data to verify or replicate published findings, but also to provide them in a form for reuse. Thus, standards maximize the value of scientific findings to the community (National Research Council, 2003).

3.3.7.1. FAIR principles

The FAIR guiding principles for research data stewardship (findability, accessibility, interoperability, and reusability) look set to become a cornerstone of research in many disciplines, and they are being incorporated into Humanities discourse on research data as well. FAIR refers to a set of guiding principles that support increasing reusability, via many different implementations. They are sometimes incorrectly referred to as a ‘standard’, allowing many different approaches to rendering data and services Findable, Accessible, Interoperable, to serve the ultimate goal: the reuse of valuable research objects (Mons, 2017).

European policies on research data are now aligned to the FAIR principles, and they are expressed in science-centric language, not using broad expressions applicable both to science and arts. This causes the arts and humanities researchers to feel marginalized and disengaged (Kotar, 2018).

3.3.7.2. Data Citation Principles

Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse. The Data Citation Principles cover purpose, function and attributes of citations. These principles recognize the dual necessity of creating citation practices that are both understandable by humans and machine-actionable. Several of these principles actually pertain to data publishing, like:

- Importance: data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.
- Access: Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.
- Persistence: Unique identifiers, and metadata describing the data, and its disposition, should persist - even beyond the lifespan of the data they describe.
- Interoperability and Flexibility: Data citation methods should be sufficiently flexible to accommodate the variant practices among communities but should not differ so much that they compromise interoperability of data citation practices across communities.

Other standards which are relevant for data publishing and reuse include DC (Dublin Core), EDM (Europeana data model) and EAD (Encoding archival description).

3.3.8. Best practices

Within the context of this pilot, we have examined projects that can serve as best practices for publishing data and building an infrastructure for advancing data sharing. There are initiatives focusing on widening the access to data through the development of digital archives that are reusable in an open access framework. Since the EC supports research infrastructure (RI) developments in the Humanities

---

12 Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/group/joint-declaration-data-citation-principles-final]
with special attention to the field of Digital Humanities, there are several projects, such as DARIAH ERIC\(^{13}\), DIGILAB\(^{14}\), KPLEX\(^{15}\) projects, that have the agenda of creating RIs, including the development of networks of facilities and resources and services offered to research communities to support their work.

These initiatives help us defining the context, elements and work process for developing and implementing a data journal. Some of the best examples of data publishing projects include HumaReC\(^{16}\), a Swiss National Foundation Project, which aims at developing a certified and continuous data publishing digital research in the Humanities (Clivaz, 2017). The established book publishing, the traditional Humanistic way of disseminating research results, has been challenged by the a completely new paradigm by transforming research through digital writing material. New methods of dissemination have surfaced in the form of videos, draft papers, social media posts, short syntheses of datasets in blogs before the research is completed and peer-reviewed (Schulthess, 2017). The project sets out to observe the changes happening in Digital Humanities research by continuously publishing their data related to the project research on an open platform: three publication formats were chosen for data dissemination: (1) a virtual research environment with ISSN where all the published material associated with the project is collected, (2) a research blog on the development of the project, and (3) an open access web-book summarizing the research in a narrative. These results will provide first-hand experience on disseminating research data through open channels of communication and on the community-based commenting and review practices.

Another example is project PARTHENOS\(^{17}\) that focuses effort more on improving and maximizing access to and the reuse of research data through the development of Humanities data template. The development of a PARTHENOS data management plan which builds on the Horizon2020 DMP template aims at addressing the domain-specific procedures and practices within the humanities, taking into consideration standards and guidelines used in data management that are relevant for PARTHENOS specific research communities. Besides standardization in the areas of documentation of primary data and sources, reference resources and procedures and protocols, tasks in the project address interoperability and semantics through defining a common semantic framework, and designing resource discovery. The ultimate goal in PARTHENOS is to define the technical development of the tools and services that are required to create the desired trans-humanities research infrastructure. The project results here provide an excellent example for building a cross-disciplinary environment to enable humanities researchers to have access to data, tools and services based on common policies, guidelines and standards (Bassett, 2017).

OPERAS\(^{18}\) provides pan-European infrastructure and services for open access to social science and humanities research requires widespread coordination and support, as well as funding from supporting countries. This can best be achieved by application to the Roadmap of the European Strategy Forum on Research Infrastructures (ESFRI) which supports the development and implementation of mature pan-European research infrastructures. Within the recently published Operas White Paper the importance of a common operational framework for digital publishing is acknowledged. Furthermore, it is emphasized that the effective implementation of common standards is highly depended upon stakeholders’ increased awareness and commitment towards more effective ways of conducting, presenting and communicating research. The OPERAS network supports and wishes to coordinate an initiative for the introduction of publishing standards, in the SSH and beyond. A sustainable approach for the introduction of operational and technical standards for electronic publishing should take into account discipline-specific standards, user needs and domain specific research and publishing workflows (Operas, 2018).

---


\(^{15}\) [https://kplex-project.eu/](https://kplex-project.eu/)

\(^{16}\) [https://humarec.org/](https://humarec.org/)

\(^{17}\) [http://www.parthenos-project.eu/](http://www.parthenos-project.eu/)

\(^{18}\) [https://operas.hypotheses.org/](https://operas.hypotheses.org/)
3.3.9. Data journal framework and workflows

The development of a data journal framework involves a description of the communication flow and a breakdown of this process into single steps. The data journal framework should include the following attributes:

- the assignment of persistent identifiers (PIDs) to datasets,
- peer review of data
- metadata information and technical check
- links to related outputs (journal articles)
- facilitation of data citation
- standards compliance
- discoverability (indexing of the data).

The RDA-WDS Publishing Data Workflows Working Group (WG) has developed a data publication process (Figure 1) which shows the double route of data publishing through repositories and data journals. These two methods of publishing are interlinked since the formal publication of data sets in in a data journal usually requires the deposit of data sets in a repository. Both ways of data sharing adhere to specific sets of standards to ensure discoverability and reuse of the datasets. However, the review process may differ. The submission to a repository is reviewed from a variety of perspectives depending on the policies and practices of the repository including formatting issues, content, metadata or other technical details and version control of the data set. The data journals are similar to the traditional research journal, thus besides the formal peer review of the validity of the data, the dissemination of the datasets is included in their communication processes. Reviewers have pre-publication access to the datasets, so it is important that the data journals and the repositories adhere to the same standards of processing data and workflow coordination (Austin, 2017).

Figure 1. Research data publication workflow.

https://doi.org/10.1007/s00799-016-0178-2

There are similarities between the peer review of an article and the peer review of data, but data review includes more issues to manage. Research datasets can be really complex, multifaceted information objects. Some datasets are constantly changing and being added to over time, so version controlling is a necessary feature added to the data sets. Review of data should involve the checking and evaluation of the methodology of the data collection, the examination of the software code used to process the data (Carpenter, 2017). Therefore, data peer review is conducted by invited domain experts external to the
data journal (e.g. RIO journals). With the rise of open or post-publication peer review, some data journals are also inviting the wider community to participate in the publication process.

The data publishing workflow should provide information about the types of data accepted for publication, the roles of involved stakeholders including authors’ guidelines and requirements of publishing, rules of the review process (options and rules of open peer review), standards used, identifiers accepted and their connection (linking of identifiers, such as ISNI, VIAF, GND with ORCID).

3.3.10. Issues to be solved: workshop results

OpenUP organized two events where the issue of data sharing was discussed in detail and the feedback from the representatives from Humanities communities shared: a workshop at the DARIAH annual event in 2018 and the OPENUP/FOSTER joined training workshop.

During the small group discussion section of the workshop the topic of data sharing and data availability was examined more in-depth. Participants were given a poster on which they could record examples of good practice, barriers and challenges to implementation, and (based on these barriers) what actions should be taken, by whom. The group rotated so that each group moved on to evaluate and validate the findings of the other groups. They had the option of adding any points they feel were not covered by the previous group. We received valuable input from researchers and publishers.

Participants listed the following challenges and barriers that hinder the uptake of data sharing and data publishing:

- some disciplines are less willing to share materials than others,
- unclear intellectual property rules and licensing,
- data ownership issues,
- technical aspects of linking research outputs,
- lack of incentives to do the extra work (reformatting, anonymizing, making datasets platform ready).

The actions needed to solve these issues recommended by the workshop participants are the following:

- raising awareness of licensing option, data ownership issues, intellectual property issues,
- developing and implementing data documentation processes,
- including steps on data curation in the regular research workflow.

3.3.11. Solutions and examples

Registered Reports, an increasingly required step in the research workflow in social psychology and in certain fields of life sciences, may contribute to open data sharing. Registering the research procedure before the actual data collection makes explicit how the research is conducted and make all the relevant materials and data available to facilitate reanalysis, reuse, and replication. A special issue of the journal Social Psychology in 2014 presented 15 articles with replications of important results in social psychology, which included Registered Reports in their research and publishing process. All these articles demonstrated that the original proposals, anonymized data, and study materials are registered and available at the Open Science Framework (OSF\(^{19}\)). Furthermore, each article earned badges acknowledging pre-registration, open data, and open materials and the badges and links to the OSF projects appear in the acknowledgments section of each article (Nosek, 2014). Thus, Registered Reports practices provide authors the workflow how to include data in the publishing procedure and the incentive to share data and materials with the research communities.

The Campus Laboratory DCA at the University of Göttingen is an interdisciplinary centre for research on innovative digital methods in the Humanities and Social Sciences with a focus on data analytics. Data on the research workflows collected from the projects and institutes belonging to the Campus Lab DCA and infrastructure shed light on Humanities data management and curation practices. Information gained within the data collection process demonstrates that

\(^{19}\) [http://osf.io/](http://osf.io/)
• the data these Humanities projects work with are greatly varied,
• the storage and long-term archiving is usually solved internally within the project either on the institute’s server, computer or the department’s cloud service,
• data sharing is managed mainly within the closed research environment e.g. through a virtual drive which can be accessed by request (www.landesgeschichte.uni-goettingen.de/chausseen).

3.3.12. Lessons Learned

These practices demonstrate the lack of standardized workflows for data curation, sharing and publishing. Humanities data management practices at the University of Göttingen demonstrate a varied picture with various degree of openness in regard to archiving and sharing data within the research groups and with external researchers. Humanities projects and departments could take advantage of the institutional repository infrastructure or the developing DARIAH data repository services where standardized data templates, workflows and added quality assurances tools could provide a more consistent view on data publishing across the different disciplines in Humanities. Implementation of standards and guidelines for managing research data would definitely support a more common view on data sharing and data availability within Humanities projects. In many cases the tools are given for data publishing (e.g. psycholinguists are using a platform for data analysis which allows the publishing of the description of the data set, a data paper in a push of a button), however the awareness around the benefits and value of sharing research data is not part of their research flow. Humanities data publishing will be more prominent as awareness is increased among researchers on data management and data discoverability issues.

3.3.13. References


Lawrence B., Jones C., Matthews B. et al. (2011). Citation and peer review of data: Moving toward formal data publication. The International Journal of Digital Curation 6 (2). doi:10.2218/ijdc.v6i2.20


3.4. Pilot 4: Transferring the research lifecycle to the web (Open Online Research)

3.4.1. Introduction

In this pilot, we addressed the question whether data analysis and data collection in qualitative research can be transferred to open online groups, in which potentially both academics and non-academics can participate. A special focus lied on mechanisms to reach out to and engage citizens in qualitative research processes. Our aim was to transfer the data-analysis and data collection parts of the research lifecycle to the web. To this end, we developed dedicated software called OpenOnlineResearch (OOR) further. Figures 1 and 2 below show two of the main pages in OOR. Figure 1 shows the landing page and one exemplary project “interpreting depression”. Figure 2 shows the two main types of activities participants can engage in now: upload data or interpret data.

In Pilot 4 we successfully tested and further developed OOR. In particular, we tested the applicability of this online solution to involve citizens in qualitative research. The goal was to gain further insight into working practices and address current challenges/gaps of open online collaboration approaches applied to qualitative research.

OOR builds on prototypes developed earlier at the University of Amsterdam. At that point, the prototypes were geared to academic participants who already had some experience with similar tools. In OpenUP we worked to make the software easier to use, which allowed to involve citizens that were not experienced in research or using the tool. More precise descriptions of the earlier steps of the OOR development can be found in the intermediate report.

In a second step, it was foreseen to engage citizens through the Coursera online course platform. The initial plan to create the first Open Online Research in Coursera, however, turned out not to be practical.
Below we detail why we changed for Coursera to the Zooniverse citizen science platform. Due to this switch, the final test round with the Zooniverse community could not be concluded during the second pilot implementation phase. However, an additional thorough small-scale test round involving our extended team has been done, and we performed a small test at the demo session at the OpenUP Final Conference\textsuperscript{24}. The final test with the Zooniverse community is scheduled for the end of 2018. The final test is meant to determine the scalability of the method. The principles of the method and tool are tested in OpenUP. Throughout the pilot we attended to diversity and inclusion. We aimed at including citizens with different national, cultural, class, race, age and gender backgrounds.

\textbf{Figure 1. Screenshot of the further developed OpenOnlineResearch tool.}

\textbf{Help make sense of the world}

OOR enables you to share, discuss and organise interpretations of real-life experiences with people all over the world.

\textbf{Help to make sense of the world}

Join one of the research projects below

Interpreting depression

Help us to find out what people all over the world mean when they talk about depression. Upload your own observations regarding depression and interpret observations of others.

Contact us

Your Name
Your Email
Subject
Message

Send message

3.4.2 Software tests in OpenUP

The tests have been reported about earlier\textsuperscript{25}. In summary: we have performed three tests with team members and non-team members (uninitiated and untrained students from different disciplines plus non-academic and academic members of our network) before 2018.

In 2018, following further improvement of the tool, we had a software test with the extended team. In this test we both assessed the workings of the functionalities, the specific needs for (a minimum of) instructions and the CMS and output capacities so far. 5 different users went from log-in to

\textsuperscript{24} \url{http://openup-h2020.eu/openup-final-conference/}

interpretation to stacking repeatedly. All steps were analysed, and potential improvements were discussed. The test revealed no major flaws.

Additionally, we have interactively demonstrated and tested the tool at the final OpenUP conference. We invited conference participants to use the interpretation features of OOR, visually observe the use of it, and asked them about usability, instructions and meaning of the interpretations given. Six participants have gone through the software and performed the main operations. It turned out that users understand the flow and meaning of the software swiftly and are able to produce meaningful content. It also turned out that we omitted a small number of instructions. The test also triggered participants to think about the applicability of the tool. For example: one participant suggested to try the tool for peer-review.

The results of the test rounds confirmed that online applications such as the OOR tool enable citizens to gather and analyse data online and openly by means of a model investing in social moderation. The test round was assisted by a team of academics and the previously mentioned tailor-made software.

3.4.3. OOR: From Coursera to Zooniverse

While we were rather advanced with the Coursera course approach, we nonetheless decided to switch from Coursera to Zooniverse. This change was driven by two considerations. First of all, Coursera is becoming more commercial which forms an obstacle to participation. UVA runs several courses on Coursera and have made some discouraging experiences related to this development. Second, the need for training in Coursera became less pronounced since the OOR software testing enabled a much easier to use design as a standalone tool. Finally, we produced a number of online tutorials at our disposition for participants who desire methodological training about OOR.

Zooniverse is the world largest citizen science platform with about 1 million users. It contains dozens of citizen science projects, largely in the area of the natural sciences, partly in behavioural science (psychology) and humanities (mainly history and translation of handwriting). The number of participants in separate Zooniverse research projects is mostly about one to three thousand. The total number of volunteer participants is roughly one million.

Since early 2018 we exchanged information with Zooniverse. They were interested the way our tool enables collaboration since current Zooniverse projects are largely based on distributed tasks rather than direct collaboration between citizens. The Zooniverse software does not enable direct collaboration yet. The current platform and projects are often about visual data, while OOR contains a methodology for textual data. Zooniverse were also interested in the interpretive approach of OOR, which is novel and still underdeveloped in the field of citizen science. OOR has developed an approach to include diverse perspectives which is different from Zooniverse and other citizen science platforms.

In short, the switch from Coursera to Zooniverse confirmed that our approach and tool are promising and opened new possibilities for OOR to reach out to citizens.

Collaboration with Zooniverse has helped us to further develop and simplified the OOR tool, since it further pressured us to design for a largely unknow and diverse group of participants. Apart from data output, most parts of the software are developed and tested.

26 https://www.zooniverse.org/
3.4.3. After OpenUP

After OpenUP and without the use of OpenUP funding, we will perform a test in Zooniverse. The test will start late 2018 and run for 1 to 4 weeks, depending on participation. We expect early large-scale participation which would mean a test period of a week.

The Zooniverse test will focus on scaling up and the pitfalls we encounter when scaling up. We have three methods to evaluate the final test.

1. Back-end data analysis: Which proportions of participants contribute observations, generate interpretations, stack interpretation, chat with each other, adjust interpretations etc?
2. Analysis of the quality: Researchers compare the contributions to existing research.
3. Survey to participants: After their contribution has ended, participants receive an invitation for a survey about: user experience, learning outcomes, experience of collaboration, empowerment. In the analysis of the survey data, specific attention will be given to gender, class and regional differences.

3.4.5. Lessons Learned

Within OpenUP, Pilot 4 demonstrates that open online interpretation of qualitative data is feasible and that yet unused parts of the research cycle can be opened to wider ranges of collaborators both within and outside academia.
The results of the testing confirmed that the Open Online Research (OOR) tool enables online collaborative interpretation. Three iterations of the tool have been tested by various groups of participants (5-40 people). A further test falls outside the reporting period but it builds on the insights from the tests done in the frame of OpenUP.

More precisely we learned that a simple tool, without the need for detailed instruction, is feasible. We also learned that the input of scientists is still needed for the formulation of sound research questions and instructions. We have also seen that online collaboration needs moderation (either technically or by humans) to settle differences. However, we found that conflicts were rare and that participants were willing to collaborate in most cases.

The outcomes of this pilot are feeding into future developments in open science in two ways. First, the collaboration with Zooniverse is continuing and might lead to either an integration of OOR methodology in Zooniverse or to a strengthening of OOR. Second, the continuity is safeguarded by new funding appointed by the University of Amsterdam. Based on the progress made within OpenUP, the UvA was willing to invest into the development of service package for OOR. While OOR is designed to be open and freely available, certain users might have more elaborate needs when using the tool. The need for services is assessed with the funding. In short, OOR is evolving beyond OpenUP.

3.5. Pilot 5: Addressing & reaching businesses and the public with research output

3.5.1. Introduction

The goal of the fifth OpenUP pilot study was to analyse and test how disseminated research results can be made more interesting, appealing, and usable for target audiences beyond the research community. In this pilot we particularly addressed dissemination to businesses and the general public.

In context of Task 4.4 we interviewed 7 science communication experts (4 male, 3 female) to define requirements and expectations by these targeted audiences. In addition, we consulted one of the community contacts of the previously involved SmarterTogether28 project, who was responsible for project communication. Based upon the feedback gathered, we created guidelines and recommendations for researchers who want to communicate their research to target audiences beyond academia.

In a second step these guidelines were tested by a research project in the Energy research area (ReFlex, a European smart grids project29). Based on the provided recommendations and guidelines, the research community re-shaped and evaluated their dissemination strategy and produced targeted dissemination content tailored to its stakeholders. Feedback was collected in an informal discussion and an interview with 3 female and one male interviewees. They were people working directly with project dissemination. The collected feedback on their experiences in applying the guidelines is reported below.

The feedback from the ReFlex project gave us very valuable input to improve the guidelines. In particular we added one additional step regarding monitoring and implementing the dissemination strategy during the project runtime. The guidelines are currently being updated addressing the feedback received and will be provided on the OpenUP Hub30 very soon.

3.5.2. Testing and feedback from ReFlex project

The main evaluation of Pilot 5 consisted of a qualitative part:

- Test the communication guidelines produced in Task 4.4 in the ReFlex project (the updated and final version of the recommendations is included in the Appendix)
- Dissemination of research content by the ReFlex project targeting businesses and the general public according to the OpenUP guidelines

28 https://www.smartertogether.at/
29 http://reflex-smartgrid.eu/
30 https://www.openuphub.eu/
In March 2018 the ReFlex project agreed to become the new cooperation partner for the evaluation phase of the fifth OpenUP pilot study. During the period from April to June 2018 four individuals from ReFlex, who were in charge of the project’s communication task, tested the guidelines in context of the dissemination activities of their project. Due to the late replacement of the SmarterTogether project, which dropped out from the cooperation early 2018, it was not possible to apply and test the guidelines more extensively.

Qualitative feedback on the ReFlex project’s experiences was gathered from three community contacts in an interview. The interview was structured by 5 main guiding questions (below represented in italic).

We gathered information about their general experience during the testing of the OpenUP guidelines as well as on the suitability of content and structure of the guidelines. Finally, we also asked their opinion on what could be improved in the current version of the guidelines. Additional feedback that was provided by one of the SmarterTogether contacts during the creation phase of the guidelines was considered as well.

**Question 1: In general, how was your experience during the testing of the OpenUP guidelines?**

In general, feedback was positive. The community contacts confirmed that the guidelines provided good support for setting up a structured dissemination and communication plan. The guidelines were helpful to assess and revise the project’s dissemination strategy. In particular the single steps and guiding questions provided a good orientation and check for shaping the dissemination strategy and make it more target audience focused. The expert tips, however, were not perceived to be that useful. Feedback suggested that in particular the expert tips for businesses are in need of improvement.

The contacts told us that they liked the structured approach: it was helpful for them even if they already had a dissemination strategy at the time of the testing. It helped them to see what needs to be considered for the overall communication strategy planning and the later steps.

Finally, the community contacts confirmed that the guidelines are useful to shape/define a communication strategy. However, they stressed that they are not so useful/not sufficient to compose the final communication message itself. They provide orientation but not a real guidance about how the final communication message or media should actually look like.

One of the community contacts reported that she used the guidelines for another project as well. With the help of the guidelines, she was able to set up a communication plan for a conference in 15 minutes. This suggests that the guidelines are easy to use and well applicable to other contexts as well.

**Question 2: Were the guidelines useful to better define your dissemination strategy towards addressing businesses and/or the general public?**

In the case of the ReFlex project, the guidelines were not that useful to better define their target audiences because these were already well defined before. In the beginning of the project they had a communication plan, which was not target group specific at first. The target groups crystallised only during the first phase of the project (the project runs for one year). However, the community contacts confirmed that the guidelines would have been useful for this purpose if they had not already undergone this process. During the first project year they posed themselves similar questions to define their target audiences, and the guidelines would have supported this process.

Another positive effect of the guidelines was that at the beginning of the project they did not compose target audience specific tweets. This changed after they used the OpenUP guidelines.

The community contacts reported that they had difficulties in matching the two main target groups in the guidelines (businesses and general public) with their own target group definitions. In the ReFlex project communication to the general public is a clear dissemination goal; this was not an issue. However, businesses are not part of their targeted audiences. Instead, they are targeting energy experts as well as local authorities/municipalities. At the beginning it was not really clear for them which recommendations to follow for those target groups. From the description that we provided, it was not clear to them which target groups we had in mind when we drafted the recommendations. From the
guidelines they understood that we divided the recommendations into expert (businesses) and non-expert audiences (general public). From the ReFlex project’s experience, however, municipalities lie between expert and non-expert target groups (i.e. semi-experts). Also, the general public is a crucial customer for those target groups (e.g. communication for cities).

Earlier feedback provided by one of the communication contacts of the SmarterTogether project also suggested that the terminology is not necessarily straightforward. Next to small and medium businesses in the Energy area and members of the general public, local authorities and municipalities were also key audiences of the project. However, municipalities were mainly stakeholders of the project, and not so much target groups for communication. Feedback suggested that it is important to differentiate between stakeholders and target groups, and that both have varying implications for the communication strategy. Following to this comment we decided to stick to the target group terminology.

Feedback from the ReFlex project contacts suggested that the information regarding general public provided in the guidelines was more conclusive. However, the businesses target audience did not fit well with the audiences targeted by the ReFlex project. Businesses are experts with a very specific background which does not necessarily fit other, not business-oriented audiences that are also targeted in European research projects. The division into these two main target groups is rather arbitrary and cannot necessarily be applied to other specific target groups. From the beginning, we did not aim to develop overarching communication guidelines. However, the testing showed that more flexible guidelines including more target groups would be useful.

We decided to keep the differentiation between businesses and general public target audiences as this was what we defined in the Description of Action. However, already the dialogues with the science communication experts suggested that there are other target groups as well, and that they do not fit in the schema that we provided.

Question 3: Are the 4 steps well-chosen and in a logical order?

As reported above, the four steps were perceived to be very useful for planning the dissemination strategy. The community contacts commented that the steps are not necessarily part of a linear process. Defining objectives, target groups and key messages are clearly interconnected steps and can be iterative.

What they suggested was missing is a final step on how to implement the guidelines during the project runtime. The guidelines did not give advice on how to plan and implement a process to make sure that dissemination is being done in parallel to the project execution. Having this in place as well as to assign responsibilities are key criteria for a successful implementation of dissemination during the project runtime. It can also be a circular process, in which the previously defined objectives, key messages etc. are being assessed and refined during the project.

Question 4: Is the format of the written guidelines and the overview table appropriate? Which one was more useful for you?

The community contacts said that they mostly used the provided overview table. The table is written similarly to a checklist, which is very useful. The table summarises the key elements very well.

The length of the guidelines was perceived as good. The written text is less well suitable to have a ordered and quick guidance. This clearly gave us the confirmation that we needed to work further on the overview table as it is the better format to work with.

In the overview table, more background information would have been useful in some cases. More information is contained in the text document; however, it is not straightforward to look it up. Some sentences in the overview table were too short and not really helpful.

The objectives were the most useful for the ReFlex project, followed by the guiding questions. The expert tips, however, can be improved. Especially for the business target group they are not very helpful (they lack clarity and are not well applicable).

Question 5: Where do you see room for improvement?
Additional feedback from the ReFlex contacts gave us some hints about content and restructuring possibilities. An input was that **trans-disciplinary questions** are also often a challenge; they are not included in the guidelines. Also, **policy makers and other target groups**, which do not fit in the provided schema are missing, respectively it was not really easy to assign other target groups to this schema.

**Practice examples** are missing in the guidelines but would be really useful.

An alternative idea to make an infographic of the guidelines that was suggested by one of the contacts is to depict a linear process of the four steps with arrows in between.

**Suggestions for the additional 5th step:** The community contacts suggest that if one of the project partners is responsible for the dissemination and communications WP and the appointed person is engaged, it works. However, during the project runtime it is not only important to define key messages. It is also crucial to define who is in charge for communicating the messages. Who communicates with whom? Is it the same person, or are there various contact persons/persons responsible for communication? This might differ in the various project phases. This should also be considered in the planning phase.

Another point is the definition of communication phases: how regularly should communication happen during the project. Continuity of communication is essential for a good project communication (sustainability). Another important point during the project execution is to reflect and assess if the communicated messages fit the audience, the medium, etc.

### 3.5.3. Impact of ReFlex project’s Twitter Activity and Reached Target Groups

The second part of the Pilot 5 evaluation consisted of a quantitative analysis of the achieved impact metrics of the project's Twitter channel and a qualitative analysis of the reached target groups. The goal was to explore if Altmetrics can be used as a meaningful indicator for assessing impact in specific stakeholder groups. In particular, we wanted to test if additional information about the reached target groups can be extracted by means of Altmetrics to answer the question if the alternative dissemination methodology applied helped making the research outputs more interesting, appealing, and re-usable.

Due to the delay caused by the drop-out of the Smarter Together community from the collaboration with OpenUP, it was not possible to analyse the impact of the composed tweets for a long enough period of time to allow meaningful conclusions. The ReFlex project only joined the pilot study in March 2018 and started testing the guidelines in April 2018. The project’s main interactive online communication channel to reach out to the general public is their Twitter channel31. During the guidelines testing period (i.e. from April to June 2018) they composed only five tweets on Twitter and had a rather modest impact. The two tweets with the most interactions had 3 re-tweets and 2-4 likes.

A comparison with the tweets sent previously to the OpenUP guidelines testing phase does not allow any conclusions about the success of the adapted communication strategy. The tweeting activity of the ReFlex project’s Twitter account started in June 2017. Only two of the 5 tweets composed during the testing phase had a somewhat greater impact compared to the other tweets. However, since the number of tweets, re-tweets and likes is still very low, it is not a relevant increase of impact. To be able to evaluate the success of the applied OpenUP guidelines in terms of increased impact, it would be necessary to observe the project over a longer period of time.

The qualitative analysis of the reached target groups was done by manually looking at the profile picture, the short text (incl. hash-tags) included in the Twitter profiles, and the history of tweets of the individual accounts from which the re-tweets and likes were made. The analysis of the Twitter profiles from the two most successful tweets shows the following distribution of accounts re-tweeting or liking the tweets:

- **Tweet from 21st of May 2018** (3 re-tweets, 4 likes)
  - 3 individuals with clear interests in research topics; 1 female, 1 male, 1 unknown
  - 2 individuals with other/not clearly attributable interests; 1 male, 1 unknown

31 [https://twitter.com/reflexsmartgrid](https://twitter.com/reflexsmartgrid)
Looking at these results it is immediately clear that it is not an easy task to determine the reached target groups. For one, the information that one can deduce from the Twitter accounts is limited and, in some cases, not evidently pointing to one target group or the other. Only one of the cases above, the real estate developer account, was evidently belonging to the businesses target group. Some accounts had clear references to research topics in their profile descriptions. We could deduce that those individuals either are researchers or citizens with particular interest in research topics. However, the other individual accounts did not contain any clear reference allowing us to assign them to any specific target group.

3.5.4. Thoughts about analysing the reached target groups via Twitter

By looking at the Twitter profiles and the tweets of individual accounts, it is not always evident to which target group the reached individual belongs. For instance, even if the accounts clearly included references to interests in research topics in the profile description text, the tweets and re-tweets from their Twitter history did also refer to other topics such as politics. Accounts of individuals can be used very personally, professionally or both. This makes it difficult to draw conclusions about the stakeholder or target group an individual belongs to.

An additional difficulty are fake accounts (in fact, one of the individual accounts was marked by Twitter as restricted due to suspicious activities) and private accounts without a public profile. Determining the reached target audience is highly depending on contextual information. If this information is not provided or restricted, it is not possible to make any deductions in terms of target groups reached.

An additional challenge is scalability. Due to the low number of tweets we did not test scalability of this approach. However, by using the Twitter developer’s API and collecting and analysing the data by means of an R script developed by DZHW it is possible to analyse the impact metrics of twitter channels over an extended period of time semi-automatically. With this method it is also possible to semi-automatically extract and analyse information about the Twitter profiles, which re-tweet and like the tweets in question and thus analyse a larger sample of Twitter profiles. However, the above listed issues remain, and it is questionable if the resulting analysis can be used as a meaningful indicator for assessing impact in specific stakeholder groups.

3.5.5. Lessons learned

Summarising we can say that the guidelines have proven to be useful for shaping/defining a communication strategy for a research project targeting these two large audiences. However, they do not give enough information and guidance for composing the final communication message as such.

Feedback from the ReFlex project gave us valuable feedback that allowed us to improve the guidelines further; in particular their suggestion to add an ulterior step on how to plan and implement a process to make sure that dissemination is being done in parallel to the project execution.

What could be re-evaluated and expanded is the chosen terminology and the defined scope and target groups (e.g. to include trans-disciplinary questions or guidance for addressing ulterior target groups). This does, however not lie within the scope of this project. For future research it would be relevant to explore other ways to structure the guidelines and their content to provide additional guidance for the points that our guidelines fail to provide substantial support.

Our pilot did not provide enough evidence about the measurability of impact at the targeted audiences by means of analysing likes and re-tweets by Twitter users. Our results suggest that it is not as straightforward to draw conclusions about the kind of target group Twitter users belong to. It would, however, be interesting to analyse this further with a larger dataset.
3.6. Pilot 6: Reflexivity of metrics on medical research and dissemination practices

3.6.1. Introduction

The goal and scope of this pilot study was to explore how biomedical research communities deal with opening up their research enterprise and how reflexive engagement with research practices at their facilities might help to develop metrics and incentives for research organization. The stance towards our partners which led us through this activity was to really focus on the needs of this biomedical community to reach their targets in the realm of open science.

For this pilot, we managed to cooperate with the Berlin Institute of Health (BIH) as a project partner. The BIH is a special facility because it combines the practical experience in clinical research facility (the Charité) with the competences in molecular biology research (at the Max Delbrück Centre for Molecular Medicine, hereafter, MDC). The BIH seeks to develop new practices, processes and guidelines, which help to create bridges between what is called laboratory research on the one, and patient oriented research on the other hand. To achieve this target, the BIH wants to spur Open Science practices at both facilities, by providing new funding instruments and better infrastructures. The BIH could thus be best understood as a kind of meta-organization which aims to serve both of its members, the aforementioned Charité and the MDC.

3.6.2. Targeting the needs of the biomedical community

As we wanted to customize our service, our first step in this pilot was to identify the needs of this community. This has been reached with members of this community via 6 meetings between April and August 2017. In these meetings we identified what the BIH is already doing in the realm of Open Science. They have established an Open Access fund for which researchers from both institutions can apply. In addition, they have also started to monitor the Open Access activities at these institutions. Second, the BIH has also established an Open Science program which selects excellent biomedical researchers that disseminate their research in an open and transparent manner aligning with the Open Science principles. Yet, for reaching the goal of actively promoting Open Science, there is still need to gain more information on the provision of open data in the respective fields of biomedicine in order to potentially develop new measures or incentives. From an Open Science perspective, the open and free provision of data is an important aspect of metrification, because currently there are less indicators and hence less opportunities to incentivize such an important part of scholarly output.

In order to deal with the topic of Open Data in biomedicine, we analysed debates in editorials of major biomedical journals (N=144). Based on this analysis and the discussion we had with members of the BIH, we found that there may be different cultures of dealing with data in the biomedical sciences. Clinical, pre-clinical, and lab-oriented research have developed different practices and stances towards the handling, governing and acknowledging research. We therefore agreed to focus on data use and data stewardship at different stages of the biomedical enterprise (clinical and pre-clinical research). Clinical, pre-clinical, and lab-oriented research have developed different practices and stances towards the handling, governing and acknowledging research. On the basis of field work and exploratory analysis, we aim at providing recommendations for specific metrics appropriate to the needs of our community. Therefore, our goal was to identify barriers, enablers and constraints in Open Data in the biomedical research field. Based on this, we developed four different criteria for the evaluation of our pilot.

1. The extent to which we identified problems related to the provision of open data in biomedicine
2. Extent to which we identified community needs for the use of open data
3. Extent to which we find field specific practices related to open data
4. Extent to which we provide input of how to govern and incentivize the use and provision of open data in a biomedical research facility
3.6.3. Engaging with research and publication practices at the BIH

In order to meet this goal, we developed a strategy which contained three elements. First, we assisted the process of monitoring the Open Data publication output in order to gain an overview about publication practices in the field and to select cases for exploratory case studies. Second, we aimed at constructing a search strategy which allows for identifying different ways of how biomedical researchers mention or link open data in publications. And, third, we aimed at conducting field studies at BIH research facilities both at the Charité and the MDC in order to explore current field specific data practices and potential institutional or social barriers for open data. Summarizing, our research design and our focus allowed us to carry out activities which are targeted and which respond to the needs of the community and which will put us in the position to derive balanced and field specific recommendations towards how novel forms of metrics can be established.

3.6.3.1. Quantitative Analysis

The BIH aimed at constructing a corpus of publications at both institutions between 2015 and 2016 in order to gain more information about how many publications provided their data openly. We assisted this activity by providing professional advice and again discussed our results with members at this institution. In this period, DZHW’s role was to support the process of corpus construction at the BIH. We have developed different search strategies (both at DZHW and at BIH) which led to quite similar results although we used different databases and search algorithms (while DZHW relied on a merged and to a large extent curated data set of WoS and Scopus, BIH relied on the data sets provided by BiomedCentral. More than 11000 documents in the biomedical realm had been retrieved. We have found, however, that the use or provision of Open Data in this corpus was difficult to retrieve, since the databases do not provide this information. In addition, also other ways towards identifying open data contributions from members of the BIH were difficult. We have tried to retrieve the information by screening the platform DataCite which is one of the most widely used platforms for data citations, in order to gain this information. We have found, however, almost no contributions from researchers of these institutions in this period on DataCite. Yet there is a significant proportion of publications that actually comply with the data criteria put forward by the Open Science community. As a closer examination of the corpus shows, however, publications use a variety of repositories and different ways of mentioning and acknowledging data, sometimes due to different requirements from the journals. Hence, there is no standard, and arguably, no common way of mentioning, providing or acknowledging data in biomedical research publications (at least at this organization). Our preliminary estimation is that almost 800 publications provide their data open, but that only a few attempts to make their data citable, and, as a consequence, more visible among the research communities. Thus, it can be said that key problems hindering the provision of open data in biomedicine is the lack of information and oversight due to missing infrastructures as well as heterogeneous and non-standardized publication practices regarding the provision of open data.

3.6.3.2. Qualitative Analysis/Search strategy

Based upon this finding and addressing the need of gaining more information about these publication practices, we aimed at developing a search strategy for identifying open data research papers in biomedicine. We used a selection of papers from the aforementioned corpus of publications in order to fulfil this task. Our main goals were to explore how scholars report about the publication of data. We identified different ways of how this information is provided; it can be either in the results section, or in the acknowledgements, but also in a special section headed “supplementary material”. There seemed to be no standard rule and provision very much hinged on the rules and regulations of the specific biomedical journals. The publications, however, used in most cases URLs to link them directly to the text, so this could be a first basis for automatizing searches in this respect. In addition, we have found different repertoires by which researchers mention their provision. Generally, we see that the term “data” could be a way to identify relevant sections of the text. The publications, however, do not use the term “open” to show that their data are available. Because of the enormous heterogeneity, it was difficult to construct a technical implementation until now. We discussed our results with technical experts and data scientists at the BIH who currently aim at probing different suggestions for the algorithm.
3.6.3.3. Further outreach

In exploring the aforementioned material in this work package, we found that the construction of a search strategy for open data might be of greater relevance also to other subjects, and for the Open Science community as a whole, which is why we tried to reach out with our activities. We are currently in discussions with other organizations to come to a solution for which eventually additional funding will be available. This work is still ongoing and is likely to continue after the ending OpenUP project.

Based on these results from the quantitative analysis of publications, we aimed at selecting cases for a qualitative analysis of data practices. To this end, we have selected different cases from both, the quantitative material as well as the recommendations of our partners at the BIH. The goal was to identify projects with different research practices and diversity of stances towards the provision of data. In order to find persons and research group leaders with different views towards the handling of data and open science, we relied on the expertise of the managing institution, our partners at the BIH. We then developed an interview guideline which was discussed and complemented by our partners at the BIH. In particular, our aim was to meet the information needs of our community and to provide input for the strategic actions at the institute, the development of specific incentives for Open Science and Open Data. We have scheduled interviews with altogether 5 different research groups, in each group with a PI and a doctoral student at least (10 interviews completed, 3 more pending). Concerning gender, one third of the interview partners were female, and two thirds were male.

3.6.4. Results from the interviews

The preliminary results of the interviews revealed the aforementioned different data cultures in biomedicine there are in fact very different cultures in biomedicine regarding the handling and the provision of data. We found rather different accounts of problems, but also different solutions to the problem of disseminating data in biomedicine. In total, we conducted 10 interviews and field observations at 5 different working groups covering different realms of biomedical research at both institutions, the Charité and the MDC.

3.6.4.1. Field specific practices

Our first finding was that the different interviewees reported in very different levels of concreteness about the usage and handling of data. These differences could be at first, attributed to the professional role of the interviewees, but the comparisons between interviewees from similar positions show that these differences remain. It could therefore be argued that the interviews indeed may indicate different stances and engagement in data. The data also showed very different levels of technical expertise in providing open data. While in the case of a molecular biologist, interview data and observation show that the researcher has put enormous effort to report and underline the need for cleaning and curating the data, other interviewees in clinical research reported they would not have to deal with such tasks in their daily work. This shows that indeed research practices and stances towards the engagement with data and their provision may be different.

On the other hand, clinical researchers contended that part of the clinical research is to deal with possible misuse of data which according to their perception has led to much greater level of awareness regarding the possible drawbacks of data provision. According to these accounts, clinical researchers are more sensitive towards the way data are presented, treated or disseminated, which is why, according to their perception, the current discussion on open data is very much influenced by positions from this community.

Interviews show that one of the possible topics for further exploration is to elaborate on what are considered useful criteria for “good data”. As the interviews reveal, these criteria might be influenced by very different value positions, such as the degree to which data are certified, as to whether are anonymized, controlled, or curated in order to discover false positives and so on. While a dedicated open data culture has emerged in some fields, particular in molecular biology, this is less established in other realms of medicine. For instance, one of the clinical researchers very much insisted on the criterion of blinding in clinical research, which according to his point of view is not sufficiently established in medicine. We find, however, almost no other criteria which were considered relevant. For another
professor specializing in molecular sciences, these criteria for quality in research are contingent, because the only thing what matters is regulations from the main funder in this community, the National Institutes of Health (NIH). Thus, criteria for quality of data should simply comply with international regulation.

3.6.4.2. Suggestions

We also identified problems and possible suggestions in the research field. One of the problems according to the interviewees is a missing expertise or guidance in disseminating and providing the data within some realms of the biomedical research community. According to the interviewees, more and tailored training would be needed. We could also see that more service and guidance is needed on how scholars can manage and process research data. Such training should be organization specific, as the provision and management of data needs to be part of the strategic decision-making repertoire of the scientists. In order to establish metrics, which incentivize Open Data, research managers in this biomedical institute should first of all provide services on how to open data which are field specific and which refer to the regulations in the biomedical field. There should be more institutional and organizational support for using services which make data sets citable or trackable. Researchers reported they would welcome more services providing specific competences of using and providing data openly. Particularly in the increasingly specializing research fields in molecular and systems biology (systems medicine), scholars need very specialized and extensive competences.

3.6.4.3. Further directions

The results of the interviews will be collected and discussed with the partners at BIH. Our main aim is to carve out the differences between the different fields, e.g. clinical research on the one hand and laboratory research (molecular biology) on the other. Referring to our questions, our main goal is to explore field specific problem perceptions and research practices which may hinder the provision of research. In addition, we also identified arguments of persons who accrue more critical stances. The main argument which has been repeatedly made is that, particularly in clinical research, the provision of data should not be promoted freely, but regulation should be clearer on who should share the data, for what purposes, and to what extent the data should be provided and shared with other researchers to prevent abuse. Currently, these problems only appear in clinical research, but it may be also relevant for laboratory research, as these more strongly orient on personalization of therapies.

3.6.5. Use case evaluation

3.6.5.1. Lessons learned

As outlined in the beginning, the goal was to identify problems and barriers in providing Open Data. Several meetings and interviews have shown that this is a relevant task for members of the biomedical research community, e.g., members of the BIH. The results show that to answer these questions, one really has to engage with the members of the community. These are our main findings:

- The pilot has shown that providing incentives for Open Data provision is difficult. There are different regulations, guidelines which need to be addressed.
- Incentives for providing Open Data need to be field specific. Even in the biomedical realm, there is an enormous variety of different data cultures, that is, different stances, ways of reflecting, handling and valuing data which make a unified framework difficult. Thus, metrics for incentives should be field specific and reflect the respective epistemic practices.
- There is a lack of regulation regarding the governance of data usage. Particularly in the realm of clinical research, there is a need of a governance framework of who can access, who can use and alter the data, for what purpose, and at what point in time.
- To a certain degree, there is still a lack of institutional and organizational support, such as guidance and technical advice in Open Data principles, technologies, and practices. Mentorships, technical or infrastructural advice might benefit the provision of Open Data.
3.6.5.2. Quantitative Evaluation

One possible way to evaluate the pilot is to quantify outreach of a given activity. This is a bit difficult to calculate, because it is not easy to demarcate the boundaries of the community in this case. In our pilot, we closely cooperated with a department within the BIH, which aims at providing services and solutions for spurring Open Science at Charité and MDC, the QUEST center. Altogether, we had so far 6 meetings with members at this institution. It is difficult to estimate the institutional outreach of this activity, but it could be quantified in the following way. Our main contacts were located at the QUEST centre at the BIH which until now has 20 employees (including fellows and related stuff). On a larger scale, the project has greater relevance for the BIH as a whole, which currently employs 70 people, including projects that aim at reaching out to policy and the wider public. One could argue, however, that the results of the project are of significance for the member institutions of the BIH, the MDC and the Charité, who currently employ more than 9000 researchers, because the BIH reports its activities to the boards of both of these institutions and employees from these institutions are eligible for funding and central services.

Table 1. Outreach of Pilot 6

<table>
<thead>
<tr>
<th>Description</th>
<th>Outreach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of direct partners at the BIH working group</td>
<td>8</td>
</tr>
<tr>
<td>Number of employees of the BIH</td>
<td>70</td>
</tr>
<tr>
<td>Number of employees eligible for funding and services from member institutions at BIH</td>
<td>9000</td>
</tr>
</tbody>
</table>

3.6.5.3. Relevance to the community

At the beginning of our pilot description, we have articulated that the main goal was to serve the community with this specific activity. Therefore, we have repeatedly attempted to collect feedback for the various steps. It can be stated that in the meeting, almost all members with which we were in direct contact agreed that the object of this activity was very useful for the community.

3.6.5.4. Field specificity

The second most important issue for our member institution was that we should focus on identifying different practices and stances which are associated with the provision of data and thus, are relevant for designing incentives in an institution that bridges very different epistemic cultures and communities. In the process of selection of our interviewees, we therefore particularly focused on diversity in this respect. In accordance with their different positions and epistemic cultures, different challenges and barriers of providing or governing the use of data have become apparent. We therefore could show that developing open data policies is not an easy task, because the different needs and restrictions should be taken into account.

3.6.5.5. Identification of problems, needs, and further recommendations:

One of the most relevant results of the case study was that we could identify major barriers in providing open data, which allow us to derive recommendations: First, we identified a lack of training and technical expertise among researchers for making data open, for instance machine readable as the FAIR principles indicate. Our results suggest that the organization would have to systematically explore the different tasks and competencies associated with the open and transparent data management. In addition, the publication analysis reveals that there are also shortcomings regarding the way data are disseminated and made visible. The lack of institutional visibility of the BIH at DataCite for instance

---

32 There are several dimensions of diversity in this respect: First of all, there is diversity of biomedical fields as we aimed to cover most different fields of research reaching from laboratory research on the hand to clinical research on the other. As mentioned earlier, these are quite different concerning stances towards data provision. Second, diversity comprises the aspect of experience, as we tried to cover different levels of experience reaching from doctoral students to professors. And third, diversity comprises the aspect of gender. Here we tried to get also female researchers as interview partners. As female professors only make up 10% of the total amount of professors, women are also in this sample underrepresented.
suggests that there is no strategy for promoting data sets at this institution which would also allow for constructing incentives.

3.7. Pilot 7: Piratical demand as a form of impact indicator and reaching unexpected audiences

3.7.1. Introduction

In 2016 Science published a short report on the usage of SciHub, a piratical scholarly journal article distribution service (Bohannon 2016). Set up by Alexandra Elbakyan, a Kazakh scientist, SciHub allows users to bypass journal publishers’ paywalls, so everyone can have access to journal articles for free. The report, based on a dataset provided by Elbakyan, offered a stunning insight into the underground circulation of scholarly knowledge (Elbakyan and Bohannon 2016). The colourful maps made it clear that it is not just developing countries that seem to struggle with access issues: high income European and North American countries are also eager pirates of scholarly articles.

In this contribution we provide insights into a closely related phenomenon: the underground circulation of scholarly books. Using a dataset provided to us by one of the administrators of a prominent shadow library in 2012 and in 2015 we mapped the both the supply of and the demand for academic monographs, textbooks and other learning material via piratical shadow libraries (Fig. 1). Our primary findings suggest that scholarly book piracy is a ubiquitous global phenomenon, with no apparent end in sight. If that is indeed true, we must ask, what might the consequences be for the status quo in scholarly publishing.

3.7.2. Shadow libraries

One of the major developments during the 2010’s in scholarly publishing was the emergence of so called shadow libraries – illegal archives of scholarly journal articles, textbooks, monographs, and other forms of academic work. Services like Gigapedia, LibGen, SciHub, Aaaaarg, monoskop, Hansi, or chomikuj provide free, unrestricted, copyright infringing access to more than a hundred million pay-walled scholarly articles, and millions of books for anyone interested. (Bodó 2018)
While the piracy of science is nothing new (Darnton 2003; Bodó 2011), online science piracy is a unique product of the conditions of global higher education system’s development in the last half century. The online science pirate libraries constitute only a relatively small part of much larger informal human and institutional networks through which scholarly works circulate. A recently published study: “Shadow Libraries - Access to Knowledge in Global Higher Education” (Karaganis 2018) gives an excellent overview of the different informal practices, which facilitate the often illicit circulation of learning materials in emerging economies. The study lays out the contexts, which contributed to the emergence of scholarly piracy: the rapid expansion of tertiary education in the world and the subsequent growth in the demand for books; the post-WWII concentration of the western academic publishing market and the access barriers set up by publishing monopolies; and the innovative solutions with which scholars try to bridge the gap between the supply of and the demand for scholarly works. (Bodó 2016)

3.7.2.1. Conditions of supply and demand in global academic publishing

In the decade between 1995 and 2005 the number of people with post-secondary education grew from 283 million to 725 million. Much of that growth originated in the developing world: India, Brazil, South Africa, Central and Eastern European Post Soviet countries. Unlike in the previous growth period, when the post-WWII Western countries increased their higher education sectors and their research capacities through huge public investments, this time the education boom took place in middle- and low-income countries and was much less reliant on public funds. In other words, in the last few decades hundreds of millions of people tried to gain access to global knowledge commons, while the infrastructural conditions for such a participation, in the form of well stocked, easily accessible libraries, were less than ideal. (Joe Karaganis 2018)

The rapid, global growth in the demand for scholarly works, and the tightening financial conditions of higher education coincided with a rapid concentration and commercialization of the Western scholarly publishing. A number of developments led to the current, highly concentrated state of scholarly publishing. The post-WWII boom in western tertiary education radically expanded the size of the scholarly publishing market (Tenopir and King 1997), which led to a series of mergers and acquisitions, where commercial publishers started to consolidate smaller, non-profit scientific publishers into a handful of strong, vertically integrated oligopolies (Larivière, Haustein, and Mongeon 2015). The same companies introduced a number of innovations such as different Science Citation Indices (Guedon
2001), which led to the development of a few highly sought after, therefore very valuable journals in many disciplines. These journals, in turn turned into powerful points of control in scholarly publishing. The publishers who control such key resources are able to charge excessive access fees despite the fact that every other input for these journals (the articles themselves, peer review, etc.) are provided by the academic community for free. Some argue that publishers now deny payment for the scholars for their work while they also deny the very same scholars to access their work, and this is the cause of the unprecedented profit levels in academic publishing (Mars and Medak 2015).

These twin developments of rapidly rising costs and rapidly rising demand coincided with the widespread availability of increasingly cheap reproduction technologies, first the photocopy machine, later digital technologies. On these technological platforms a plethora of practices evolved that tried to bridge the gap between supply and demand. The Open Access initiatives in the 1990’s created the standards of green and gold open access, and of self-archiving (Okerson and O'Donnell 1995; Suber 2013; Suber and Darnton 2016). The rapid proliferation of copy-shops on and around the campuses provided copies at the marginal cost of physical reproduction campuses (Bollag 2004; Chon 2010; GARWE 2014; Kaser 1969; Lin-Liu 2004; Lloyd 2004; Overland 2004; Rens, Prabhala, and Kawooya 2006). Parallel markets for scholarly works emerged through the re-importation of learning materials and textbooks from low-income countries where they were sold on a discount into high income countries (Gautam 2014). And finally, online shadow libraries emerged, which digitized, archived and distributed books online.

### 3.7.2.2. The emergence of online shadow libraries

Personal computers and digital networks provide and ideal environment for the creation, accumulation and circulation of texts in digital form. One of the earliest online projects is in fact a digital library from 1971, the Gutenberg Project which set out to digitize public domain cultural heritage (Hart 2006). Copyright issues were able to limit bottom-up digitization and sharing of copyrighted works for a while, however, when copyright wars turned the legal struggle around Intellectual Property protection into war on moral values, copyright infringing knowledge sharing practices in academia started to see IP infringement as a morally justifiable act of resistance against unjust publishing monopolies (Samuelson 2013; Swartz 2008; Barok et al. 2015).

In 2008, Aaron Swartz, a US activist published his Guerrilla Open Access Manifesto, in which he called for the liberation of pay-walled scholarly content in solidarity with scholars with no access (Swartz 2008). Around the same time various text archives, which were digitized and compiled by Russian research institutions in the early 2000’s, and which up to that moment circulated on DVDs and ftp servers started to consolidate and appear as the online shadow library Library Genesis (Bodó 2018).

These libraries are born through the cooperation of people like Swartz and Alexandra Elbakyan, who gave a name and a face to the countless anonymous individuals who maintain the global shadow library ecosystem. These two known individuals also hint at how wide the coalition of scholars around the globe needs to be in order for these shadow libraries to operate. Swartz represents the privileged Western scholars, the insiders, who have access to almost everything through their first class, well-endowed academic libraries. Some of these scholars, like Swartz, recognise their privileged position, and decided to show solidarity with others through the acts of sharing their knowledge, by sharing their digital access opportunities. They are the ones who smuggle the knowledge out from behind the paywalls. On the other end, we find Elbakyan, who represents scholars at the very peripheries of privilege, wealth and access. Such scholars are outsiders. They are the ones at the wrong side of the access paywalls. They tend to live in countries, like Russia, which have rich histories of highly efficient clandestine knowledge distribution networks to circumvent access obstacles, evade enforcement, and build influential underground knowledge repositories under hostile circumstances (Bodó 2014). They put their historic experiences into use and build and run the infrastructures necessary to archive and distribute the copyright infringing content.

Through the collaboration of scholars at the centre and at the periphery powerful shadow libraries now facilitate a historically unprecedented transfer of knowledge across the globe.
3.7.3. Data

This report is based on a number of data sources to analyse the supply and demand of pirated scholarly publications. The analysis of supply is based on the catalogue of Library Genesis. The shadow library publishes its catalogue with basic bibliographic metadata as daily database dumps. Usage data is based on to two sets of access log data provided to us by the administrators of one of the mirror services that distribute the titles in Library Genesis. The datasets were detailed enough to link the download of catalogue items to geographic locations. To conduct extra research into legal availability we occasionally queried other data sources, such as amazon.com from price and legal availability, and worldcat for library availability.

3.7.4. Results

In 2012 the Library Genesis catalogue contained 836,479 records. Three years later, in 2015, the catalog almost doubled to 1,317,424 records, and by the time of writing in 2018, Library Genesis hosts more than 2,237,940 documents, almost all scholarly publications. In addition, there is an extensive collection of literary works, comics, and of course the 100 million journal articles archived through the SciHub.

There are multiple websites with make this catalogue available. Though the exact number is fluctuating as a result of various legal, technical, organizational and financial troubles, at least one service has been consistently online in the last 8 years. The anonymous administrators of this single mirror provided us with usage logs for 2 months in 2012 and for four months in 2014/2015. In these three years the traffic on this single server tripled, from an average of 41,000 downloads / day in 2012 to an average of 120,000 downloads / day in 2015.

![Figure 2. The growth of LibGen usage between 2012 and 2015.](image)

Figure 2 shows the growth of daily downloads around the world. Post-Soviet republics, which in 2012 were heavy users of this particular mirror seem to have migrated to other services, and the traffic from these countries declined. Countries and regions, that account for the bulk of the usage (US, India, China, Europe) show average growth. On the other hand, there have been a staggering growth in Latin America, which in 2012 was hardly using (this particular mirror of) Library Genesis at all, but by 2015 they discovered LibGen, and became one of the most intensive users of the library.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe &amp; Central Asia</td>
<td>28.81</td>
<td>42.87</td>
<td>149%</td>
</tr>
<tr>
<td>Middle East &amp; North Africa</td>
<td>8.41</td>
<td>24.31</td>
<td>289%</td>
</tr>
<tr>
<td>South Asia</td>
<td>1.22</td>
<td>6.38</td>
<td>523%</td>
</tr>
<tr>
<td>Latin America &amp; Caribbean</td>
<td>0.06</td>
<td>11.68</td>
<td>19467%</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>0.28</td>
<td>3.98</td>
<td>1421%</td>
</tr>
<tr>
<td>East Asia &amp; Pacific</td>
<td>2.13</td>
<td>7.4</td>
<td>347%</td>
</tr>
<tr>
<td>North America</td>
<td>10.47</td>
<td>42.76</td>
<td>408%</td>
</tr>
</tbody>
</table>

From the table above, it is also apparent that the biggest per capita users are the high income North American and European countries.

**Figure 3. North American download locations**

In fact, just a handful of countries, the United States (11.66%), India (8.58%), Germany (5.23%), the UK (4.10%), Iran (3.68%), China (3.67%), Italy (3.30%), Canada (2.36%), Indonesia (2.29%), Spain (2.28%), Turkey (2.24%), and Brazil (2.11%) account for more than half of all the downloads.
3.7.5. What is being downloaded?

The global top list of downloaded books is an interesting mix of sex, popular mainstream literature, quantum mechanics, English grammar, feminist, and post-colonial studies, and machine learning.

### Table 2. Global top list of downloaded books via LibGen

<table>
<thead>
<tr>
<th>Total downloads</th>
<th>Author</th>
<th>Title</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>2927</td>
<td>Donna Evans, Sex Positions</td>
<td>Sex Positions Illustrated: 101 Hot Positions You Can Do Right Now</td>
<td>2012</td>
</tr>
<tr>
<td>2902</td>
<td>E L James</td>
<td>Fifty Shades of Grey</td>
<td>2012</td>
</tr>
<tr>
<td>2299</td>
<td>David J. Griffiths</td>
<td>Introduction to Quantum Mechanics</td>
<td>2004</td>
</tr>
<tr>
<td>2175</td>
<td>James Dashner</td>
<td>The Scorch Trials (Maze Runner Trilogy, Book 2)</td>
<td>2010</td>
</tr>
<tr>
<td>2160</td>
<td>James Dashner</td>
<td>The Maze Runner (Maze Runner Trilogy, Book 1)</td>
<td>2009</td>
</tr>
<tr>
<td>1749</td>
<td>Ania Loomba</td>
<td>Colonialism Postcolonialism (New Critical Idiom)</td>
<td>1998</td>
</tr>
<tr>
<td>1574</td>
<td>Randall Munroe</td>
<td>What If?: Serious Scientific Answers to Absurd Hypothetical Questions</td>
<td>2014</td>
</tr>
<tr>
<td>1546</td>
<td>Patricia Hill Collins</td>
<td>Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment</td>
<td>1999</td>
</tr>
<tr>
<td>1391</td>
<td>Gillian Flynn</td>
<td>Gone Girl [Hardcover]</td>
<td>2012</td>
</tr>
<tr>
<td>1367</td>
<td>James Dashner</td>
<td>The Death Cure</td>
<td>2011</td>
</tr>
<tr>
<td>1228</td>
<td>John Storey</td>
<td>What Is Cultural Studies?: A Reader</td>
<td>2009</td>
</tr>
<tr>
<td>1188</td>
<td>Gillian Flynn</td>
<td>Dark Places</td>
<td>2009</td>
</tr>
<tr>
<td>1144</td>
<td>C. Balanis</td>
<td>Antenna Theory - Analysis and Design</td>
<td>1997</td>
</tr>
<tr>
<td>1139</td>
<td>Veronica Roth</td>
<td>Allegiant</td>
<td>2013</td>
</tr>
<tr>
<td>1101</td>
<td>Kevin P. Murphy</td>
<td>Machine Learning: A Probabilistic Perspective</td>
<td>2012</td>
</tr>
<tr>
<td>1099</td>
<td>Paulo Coelho</td>
<td>The Alchemist</td>
<td>1993</td>
</tr>
</tbody>
</table>

It would be hard to read this list as an approximation of a global scientific canon. Compared the size of the catalogue, and the number of downloads, the titles on these lists represent only a fraction of all the traffic. Altogether 760,868 different documents were downloaded over a 4.5-month period in 2015, generating more than 16 million downloads. We can get a better view of the overall usage of the library
if we map the top 10 Dewey content classes in terms of the share of all downloads and the share of all (downloaded) titles.

**Figure 5. The utilization of the shadow library by subject matter**

![Bar chart showing the utilization of the shadow library by subject matter](chart.png)

Figure 5 gives an insight into the subject matter composition of the supply (titles in the LibGen catalogue), and the demand (in terms of the unique downloaded titles, and all downloads). We could not classify more than half of the library, as titles do not carry a Dewey identifier, because for example they were published in Russian. Though such titles make up the largest share of both the supply (57.29% of the catalogue) and the demand (28.64% of downloads), we excluded them from Figure 5. The remaining ten top Dewey content categories suggest a strong science and technology focus of the library, since these two categories enjoy the highest demand, and also since works have the highest download volume per title (26.8). Social sciences, on the other hand see the second lowest (13) download per title, while this section is big, both in terms of supply, and in terms of the number of titles downloaded.

### 3.7.6. Implications

What kind of impact shadow libraries may have on the current system of scholarly publishing? The review of the music- and audio-visual industries’ last two decades of fighting online piracy may help us answer this question. The entertainment industries spent two futile decades trying to find first technological, later legal solutions to the piracy of their products. Digital Rights Management (DRM) and copy protection technologies did not work, neither did the lawsuits against their customers. File sharing technologies rapidly evolved to eliminate all potential technical points of control, and decentralized themselves beyond reach (Bently et al. 2010, Johns 2010, Karaganis 2011, Patry 2009). The Pirate Bay is still up and running, despite countless lawsuits, imprisoned founders, and EU-wide blocking injunctions.

It seems that the scholarly publishing industry understood that it is close to impossible to efficiently fight scholarly piracy. Gigapedia, the predecessor of LibGen was relatively easy to shut down, as it relied on a centralized database, and a centralized document repository. LibGen and SciHub are much more difficult to eliminate, as they are both radically decentralized, and already exist in multiple copies all over the internet. That might also explain why there is only one court case against these services. A New York court issued an injunction against both sites, forfeited their domain names and ruled damages
against the administrators, but none of these measures had any practical effect. In the case of journal articles, publishers force academic libraries through which paywalled articles leak into SciHub to focus more of their resources on enforcement, but that did not stem the outflow of materials either (Bodó 2016). As it stands now, academic piracy seems to be an unstoppable force.

Under such conditions academic publishers have to ask themselves if the copyright and exclusivity-based business models are sustainable. For a number of reasons, the answer might still be in the affirmative. Both the US and the EU has mandated open access publishing for its publicly funded research, creating a lucrative revenue stream for publishers in the form of article processing fees which are not threatened by piracy. The fact that the scholarly pirates (the scholars themselves) are not those who must pay for the materials (the ones paying are the academic institutions, libraries, in some cases government agencies) may ultimately mitigate the negative effects of piracy, where illegal consumption substitutes sales. One illegally downloaded scholarly monograph, already priced for the library market does not diminish sales to individuals but may generate a purchase by the library at the request of the researcher who had a free sample copy through the shadow library. The net effect may well be positive for publishers.

On the other hand, in recent years academic publishing accomplished a radical shift towards data-centred business models. In the last decades academia underwent a substantial degree of quantification, where not just citations, but all other aspects of scholarly work became measurable and consequently thoroughly measured. The number and quality of publications, media appearances, funding, scholarly and societal impact, teaching, supervising, academic service, peer review, editorial work, retweets and blog post mentions can all be measured. In turn, funders as well as academic institutions are increasingly using such data to base their funding, hiring, and promotion decisions. Almost all major publisher recognised the potential of this data-based market and invested heavily in software and services which facilitate the circulation of materials related to scholarly work, and as such generate data on this circulation.

Figure 6. Kramer, Bianca, and Jeroen Bosman. "101 innovations in scholarly communication: The changing research workflow.”

Kramer and Bosman's (2015) overview of the ecosystem offers a snapshot in that process. In addition to the portfolio above, since 2015, Elsevier also acquired bepress (research showcase service), SSRN (Social Science repository), Plum (altmetrics, analytics), Hivebench (Lab data management) and Newsflow (media monitoring).

These tools do not require paywalled content. In fact, they all thrive in environments where there are no artificial technical or legal boundaries in front of the accessibility, circulation, consumption of

---

content. The most important impact of music- and audiovisual piracy was that it forced business innovations, which ultimately produced the flat-rate "all-you-can-eat" subscription services, which grew to dominate the online music and movie markets (Aguilar and Waldfogel 2015, Danaher et al., 2010, 2013). It is hard to tell whether shadow libraries had a similar influence on the aforementioned shift in academic publishing from the access-controlled-content based business models to the data-based models. Ultimately, it is not really important either. What matters is that the access model, which green repositories and shadow librarians promoted faces few obstacles to become the dominant mode of open access. The developments related to the recent negotiations between Elsevier and library consortia in Sweden and Germany suggest that publishers’ business models are met with more and more resistance from libraries (Kwon 2018). As Günter Ziegler, a member of the German DEAL library consortium’s negotiating team put it ”Most papers are now freely available somewhere on the Internet, or else you might choose to work with preprint versions,” and this gives German researchers the upper hand in the negotiations (Schiermeier 2018). Together with personal archives, preprint, institutional, green and gold open access repositories, shadow libraries constitute a vast, system of parallel access. Even if publishers could curb one component, the shadow libraries, they cannot control all the free and open access alternatives. We interpret the increased willingness of academic library consortia to suspend their agreements with academic publishers in order to negotiate better fees and more just terms as an implicit understanding of the power of these parallel access networks, in which shadow libraries play a major role.

Ultimately there are major consequences of the increasing shadow library use on the current systems of producing and interpreting academic indicators. Our pilot shows that shadow libraries are now an integral part of the systems of scholarly communication. They are part of the everyday routine of scholars of both the developed and the developing countries. However, there is no reliable, systematic insight into the use of these resources. Consequently, our academic indicators only give an incomplete and biased picture on the circulation of scholarly works. The copyright-infringing nature of shadow libraries only allows ad-hoc and fragmented insight into the circulation of works through them. Their modus operandi, on the other hand, is certain to introduce an unknown level of bias to our currently accepted set of indicators. For example, since SciHub uses leaked/shared academic credentials to provide access to pay-walled materials, the traffic as measured at the point of access, at the library through which the unauthorized access takes place will not provide an accurate picture of who uses the library resources and for what reasons. Since it is not reliably known to what extent SciHub serves subsequent requests to an article from its own archive as opposed to getting it again through a library, its impact on library usage metrics is also unknown. It is certain however, that any newly published article behind a paywall is at least once requested through a library, and that inflates library usage statistics.

On the other hand, the high usage numbers from developed countries suggest that at least some of the shadow library traffic is generated by users who otherwise could have had legal access through their institutions. That both applies to articles for which users go to SciHub rather than their own institutional repositories, and books for which users visit LibGen rather than their library print or e-book collections. Ours statistical models (not reported here) seem to suggest that in North America and Western Europe we cannot explain the high usage with serious access limitations. Instead we suspect that in these territories the convenient one-click access shadow libraries provide to full digital copies plays a role. This of course means that the official usage statistics of those resources that are also available through the shadow libraries will be underreporting the actual demand for key library resources.

Shadow libraries do not just introduce noise into the current indicators that measure the circulation and use of scholarly resources. Given their size, the intensity and growth of their use, the omission of the traffic through these libraries threatens to falsify these indicators.

3.7.7. References


Dušan Barok, Berry Josephine, Bodó Balázs et al. (2015). “In Solidarity with Library Genesis and Sci-


4. Appendix

4.1. Recommendations and guidelines for communicating research to businesses and the public

Updated table with an overview of the recommendations and guidelines for communicating research to businesses and the general public (see Table 1). The update is based on the draft guidelines\(^{34}\) and the comments resulting from the test phase done by the ReFlex\(^{35}\) project in context of OpenUP’s fifth pilot study.

Table 1. Draft recommendations for reaching businesses and the public: Overview table (Version 2)

<table>
<thead>
<tr>
<th>Steps</th>
<th>Guiding questions</th>
<th>Expert Tips</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Define dissemination &amp; communication objectives</td>
<td><strong>Think about your goals.</strong> What do you want to achieve by targeting the specific audience?</td>
<td>Involve business partners in a research project and/or collaboration. Get the business perspective to gain new insights for further research. Have a prototype developed in a real-life context. Commercialize your results. Get feedback. Spark debate. Influence behaviours. Invite collaboration. Influence decision-makers.</td>
</tr>
<tr>
<td></td>
<td><strong>Think about what your audience should be able to do with your information.</strong> What is the purpose of the communication to the specific target group?</td>
<td>Achieve/ maintain the pole position in market based on scientific validated results. Drive innovation in both products and services and increase quality using new technologies. Make products that are safe, appealing and competitive (cost reduction). Validate research done in-house. Find new R&amp;D collaborations. How do you want the information to be used? Do you want to engage research participants (via Citizen Science)?</td>
</tr>
<tr>
<td>2. Define target audience(s)</td>
<td><strong>Think about who exactly you are trying to reach.</strong> Who is your main target audience? Break down and define your target group(s).</td>
<td>List possibly interested partners, e.g. Large/medium companies with and without R&amp;D lab, startups, business sector-company association, Creative industry, NGOs, CSOs, etc. The general public is a very large group, having a specific sub-group in mind makes arising interest easier. Build a picture of your audiences, their motivations and experiences.</td>
</tr>
<tr>
<td></td>
<td><strong>Get to know your target audience, their needs and expectations of the research outcomes, as well as their preferred communication channels</strong></td>
<td>Keep businesses informed on your research project and update them on latest research developments. Clarify the IPR beforehand: make sure that relevant and sensitive results are only available for the company. Can you organise a survey or focus groups with your target audience to better understand them? Understanding will help you engage.</td>
</tr>
</tbody>
</table>

---

\(^{34}\) First published in Vignoli & Rörden 2018, Deliverable D4.2. Role description: Dissemination to businesses and the public, pp. 19-21

\(^{35}\) [http://reflex-smartgrid.eu/](http://reflex-smartgrid.eu/)
### 3. Define key message(s)

<table>
<thead>
<tr>
<th>Align your key message with what the targeted audiences expect. Make sure that the key messages and information that you provide are relevant for the targeted audience.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Think about “What can a particular business get out of your research”. Select just one or two key aspects that are the easiest to showcase and grasp.</td>
</tr>
<tr>
<td>“How does science solve problems that our society faces?” Also: people are curious and want to learn something new.</td>
</tr>
<tr>
<td>Explicitly include and address the targeted audience in the key message. <strong>Start with the knowledge base that they already have by involving their world in the story.</strong></td>
</tr>
<tr>
<td>Identify the specific needs of the targeted business and design your key messages by addressing how your research can be beneficial to the specifics of their field.</td>
</tr>
<tr>
<td>Relate to the everyday lives of your target audience.</td>
</tr>
</tbody>
</table>

### 4. Plan your dissemination & communication strategy

<table>
<thead>
<tr>
<th>Choose media, format and dissemination strategy that fit your communication objectives. Structure and prepare your dissemination material.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use different media channels to transmit your message: social media, information sheets and leaflets or advertisements.</td>
</tr>
<tr>
<td>Participate in public debates and go with the trend (social media, visual online resources, animations). Explain through the big picture; use “sensational details” to catch attention. Tell a story that is not too technical.</td>
</tr>
<tr>
<td>Follow up with your stakeholders. <strong>You can activate your target audiences and enable them to become active or do something with the content that you provide.</strong></td>
</tr>
<tr>
<td>Use face-to-face events, meetings at your venue and networking events to get in touch with the businesses.</td>
</tr>
<tr>
<td>Name contact persons. Participate in interactive, public events. Remember that laymen can give valuable insights.</td>
</tr>
</tbody>
</table>

### 5. Implement and verify the dissemination process

<table>
<thead>
<tr>
<th>Define who is in charge for communication and dissemination. Calibrate timing and assure sustainability.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is there a single, easily reachable reference person for any question from business partners?</td>
</tr>
<tr>
<td>Is your communication media continuously feeding news and receiving comments? Are new suggestions coming in?</td>
</tr>
<tr>
<td><strong>Assess the effectiveness of your strategy.</strong> Check if the chosen medium fits the audience.</td>
</tr>
<tr>
<td>Are you acquiring more business partners? Do they propose new projects in future?</td>
</tr>
<tr>
<td>Does your project originate more buzz in the social media? Are people asking questions and suggesting interesting new directions to your research?</td>
</tr>
</tbody>
</table>