Policy RECommendations for Open access to research Data in Europe

Minutes of the Workshop: Perspectives in understanding open access to research data – infrastructure and technology challenges

Part of the 10th Plenary Session of the Group on Earth Observations (GEO-X) & 2014 Ministerial Summit

14 January 2014, 9:00-17:00, Room 7 (Level 2)
Centre International de Conférences de Genève (CICG)
17, rue de Varembé, Geneva, Switzerland, www.cicg.ch

The workshop is part of the EU FP7 RECODE project (Grant no. 321463; http://recodeproject.eu/), which is addressing the drivers and barriers in developing Open Data Access, to identify a series of targeted and over-arching policy recommendations for Open Access to European research data, based on existing good practice.

The workshop was quite well attended with around thirty participants from 14 countries, including representatives from the RECODE case studies (Archaeology, particle physics, earth sciences, health and bioengineering).

Context setting presentations:[1]
Starting at 9:30, after participants were invited to introduce themselves, Kush Wadhwa, Trilateral Research, RECODE Project Coordinator, introduced the RECODE project, including its main objectives, partners and key outcomes. Afterwards, Lorenzo Bigagli, CNR, WP2 coordinator, introduced the content and purpose of the workshop.

Michel Schouppe, a representative of the Directorate General Research & Innovation within the European Commission, discussed the EC perspective on Open Access (OA). He described how following a commitment by the G8 Ministers of Research and Technological Development in June, Horizon 2020 promotes a specific pilot on Open Research Data Access (ORDA). Participating projects will be required to produce a Data Management Plan (DMP), requiring among other things, to identify a repository

[1] Copies of the workshops presentations can be found on the RECODE website.

to publish data for OA. The EC is not imposing any particular repository, though community repositories are encouraged.

In relation to the forthcoming Horizon 2020 funding stream and GEOSS specifically, Schouppe linked Horizon 2020 to GEOSS: Societal Challenge 5 (Climate action, etc.), which comprises two calls. He mentioned that GEOSS is a possible candidate repository for ORDA, and that the project will be asked to publish data as GEOSS Data-CORE (Collection of Open Resources for Everyone). He highlighted the GEOSS Data Sharing Working Group, structured in 4 sub-groups, and the GEOSS Data Sharing Principles. Within GEOSS, OA is promoted, but the programme also welcomes non-OA data, e.g. limited by national policies.

The audience raised a number of questions, including especially one about long-term preservation sustainability, as the H2020 Pilot only covers a project's lifetime. Acknowledging this is an open issue, Schouppe suggested that community repositories (and/or repositories of repositories) may be viable solutions.

After a short break, Stefano Nativi, Belmont Forum, introduced the International Group of Funding Agencies for Global Change Research (IGFA) and its inter-disciplinary (natural and social sciences) Belmont Forum initiative (BF). Funding resources for open access are available (upon application by at least three BF members), and twelve projects have already been funded. New calls will be released in April 2014, including calls for e-Infrastructure and Data Management.

Given the many efforts to coordinate global data management (e.g. GEOSS, EarthCube, ESFRI, UNEP, Eye on Earth, FutureEarth, RDA), the BF decided to be a hub to address open issues. At the last meeting in October 2013, six hot topics were identified, including work packages of the Belmont Forum Collaborative Research Action on e-Infrastructure and Data Management on Data Sharing (WP4) and Open Data (WP5). Nativi advocates career incentives to promote data sharing and the need to push standards (although standards are sometimes too numerous to be really effective), and argues that the challenge of non-authoritative research data (e.g. crowdsourcing) is a critical issue for the near future.

RECODE research findings:
Lorenzo Bigagli reported on the key findings from RECODE WP2 survey (questionnaire and literature review) and case studies interviews on the existing technological barriers, solutions and best practice for Open Research Data Access. Technical barriers are not related to the will or permission to share resources, but to the capability of doing it. In other words, stakeholders may be willing and authorized to share resources, but unable to do it.

According to Bigagli, the online questionnaire received around 50 responses, mostly disseminators (62%) and producers (29%; mostly natural and computer science). Among the main results are the following: The decision on what to preserve is seen from a collective, shared decision approach. End users, disciplinary association, peer reviewers should decide what to purge. Instead, the decision of what to publish online/offline is left to the producers. Primary responsibility for storing European research data is assigned to specific repositories (cf. maybe a new profession is emerging). As for the factors with the greatest impact with the uptake of ORDA: data documentation is prevalent for both producers (60%) and disseminators (80%); both consider also data preservation and data quality (46%). Then, producers focus on heterogeneity of data formats (46%; possibly their primary issue when generating data). Disseminators focus on application interoperability (43%; big issue when you want to serve users). Noticeably, bandwidth is considered not important (overlooking Model Web, workflows, etc.; seems to imply a classic batch/offline model of computation. Other results stressed the importance of metadata, and an overall half-full/half-empty glass of ORDA adoption and experience.

The interviews with RECODE case studies have highlighted that, in general, there is experience with OA publications, but not with data publications or data preservation. Data management plans are being developed, but still at an early stage. Solutions for data management and preservation are neither common nor centralized. Metadata are considered crucial to enable retrieval, re-use and preservation of research data. Overall, financial and legal barriers are considered higher priority then technical ones.

Starting from the issues identified in the context of Big Data, Bigagli presented an overview of the Infrastructure and Technology challenges (heterogeneity, sustainability, volume, quality, security) per each stakeholder category (creator, disseminator/curator, funder, end user).

Following the presentation key questions were presented to the audience, which were also brought forward into the open discussions in the afternoon:

- Where should open research data be stored and made accessible?
- How can we mitigate the technological barriers to Open Access to research Data?
- How can we cope with technological sustainability and obsolescence?
- What emerging technologies could be optimized to promote ease of deposit and retrieval of research data?

Max Craglia, JRC, underlined that data include Public Sector Information (PSI), especially for applications like the ones targeted by the INSPIRE Directive, and PSI is very volatile. The issue of reproducibility is prominent, as data are not there anymore. Provenance and persistence should be considered big challenges.

Sally Wyatt, eHumanities, asked how we deal with multi-disciplinarity, especially with regards to no-context data, such as crowdsourcing data. In the following discussion, Max Craglia commented that the EuroGEOSS project addressed this problem by documenting the semantics of data.

Anusuriya Devaraju, Forschungszentrum Jülich, advocated the need for an infrastructure allowing different access levels, as quality is defined by different actors at different levels: scientist typically focus on raw data, whereas peer reviewer may be more interested in products.

Eric Kansa, Open Context, stressed the need for more experience before conclusions are drawn. There should be less bureaucracy, to effectively support an iterative approach.

Thomas Severiens, ISN Oldenburg, sees data citation as a way to protect the investment, which should be supported by appropriate data citation policies.

Simon Hodson, CODATA, mentioned the PREPARDE project work on data publishing workflow and peer review. The equivalent of Crossref would be needed for data. He also mentioned Dryad and Xenodo on data citation principles.

Afternoon session:
After the lunch break, Lixin Wu, Beijing Normal Univ., presented a draft GEO White Paper on the Global Spatial Reference Frame for GEOSS and Spatial Big Data, supporting integrated geo-referencing for the whole Earth System, including atmosphere, surface and the crust. The open discussions in the afternoon followed this presentation.

Jeroen Sondervan, Amsterdam University Press, moderated the first plenary open discussion on: Effectiveness, gaps, and practical significance of the existing technical solutions for Op en Research Data Access. The purpose of the discussion was to get an inventory of the issues and challenges encountered by the attendees. Key elements from the discussion follow:

In the first place, the issue of curation is an important aspect when storing, preserving and publish data. It is crucial to have extra information about the data so it will be usable in the long-term. The digital life cycle of data/information is something that needs to be clear. One of the participants stated that first comes the data policy then comes the data management plan.

We have to take into account that there are multiple forms of data and therefore multiple perspectives of what research data is. That means that besides having a set of tools for generic technical solutions it will be necessary to have tailor-made solutions. Also important is the possibility of co-production. Datasets are – most of the time – the work of more than one individual. It is important to have technical solutions in order to work jointly. The development, structuring and storing of data should be well documented to make it useful and accessible in the future. Without proper documentation, it’s difficult to replicate and validate the research based on the available data.

A participant says that we should approach the challenges from a bottom-up perspective and create an informal and iterative approach to localize issues with interoperability and sustainability. We should start small pilots with software and see how it actually works. Identify what the issues and challenges are and then formulate policy guidelines based on the experiences in the field.

On quality issues, quality reviews of the data are considered essential. Tools are needed to conduct quality assessments of the data (e.g., peer review of data, software control). To get a clear idea of the quality and the usability of data you could think of user feedback tools. Someone mentioned the use of ISO standards for storing, preserving and publishing datasets. Solutions for distributed online source code management, such as Git, are reportedly being used for metadata management in some contexts (with GitHub offering free online tools), but there are not fixed standards yet, as is the case with online publications. Another aspect of quality is the way we can measure the usage of data. Data citation will become important more and more. There are no standards available yet. A shared thought is that quality should be addressed by means of provenance information, i.e. describing the methodology used for producing and processing the data. Articles that describe the data should/may be included in the provenance metadata.

Metadata is maybe one of the most important things to take into account. Without proper metadata standards and tools, the visibility and accessibility of data is more difficult, if not impossible. There are several initiatives for metadata in effect (e.g. Dublin-core working groups / Research Data – RDA Vocabulary standards). Not every dataset is getting a DOI at the moment. For online publications it has more or less become an industry standard to use DOIs. During the workshop everyone agrees that this should be the case for datasets as well. Someone proposes that the use of DOI should be mandatory. Others comment that any other mechanism for persistent identification (e.g. URNs) would be equally effective.

In the software/hardware aspect of the infrastructure for Open Access to research data, software being used to create datasets should ideally be open source. The community will ensure that issues with interoperability and accessibility will be dealt with. This cannot be assured with third party software. An important factor for researchers is that there needs to be a full integration of repositories and libraries. Libraries will be, or should be, the facilitators for data access and management plans.

Discussing security aspects, it is noted that the level of data has an influence on security, and should be captured in metadata.

Heterogeneity and de-contextualisation of data (cf. crowdsourcing) is a potential challenge for designing and enforcing security policies. This also impacts privacy aspects, especially for Public Sector Information. Human activities in general should be protected, for example with appropriate Data Management authorization policies for data access. Some participants also suggested a High Security Data Label.

Finally, the discussion noted that gaps in licensing reflect on technological issues. License information should also be captured in metadata, and that systems for data management should allow for multi-level access to data.

Bridgette Wessels, University of Sheffield, moderated the second plenary open discussion on: Recommendations on how to increase the effectiveness of the current technological baseline in supporting Open Research Data Access.

A first discussion addressed the technological environment for open access to data and raised a key point, which is that any technological infrastructure has to support the quality of the data in ensuring that it is open in a responsible way and that it can be re-used appropriately.

Security of data is important and most of the discussants feel that many of the security issues could be addressed fairly easily because the technology already exists. There is however some discussion about whether the current baseline for security is enough, and there need to be further developments. The groups also notes some new challenges in terms of privacy issues and licensing that would need further exploration.

It is thought that there are possibilities at the technological level but that there would need to be changes at the policy level, which might not be easy to develop.

The issue of quality was discussed further because it was identified as a complex topic. Points raised include that quality is potentially different for different consumers and that there are different economic multipliers. Furthermore, the heterogeneity of the data and data process makes is very difficult to choose standards. Variables include different data practices, workflows, and different ontologies. Some of the solutions for the issue of quality include developing a use model to find ways of describing data, which could then be implemented. The implementation would identify potential uses, which could combine the different aspects of making data open. There was some discussion about whether or not the researchers who produced the data could come up with a quality assessment or whether other researchers as peer reviewers could do that. However, participants felt that there is no such thing as a 'dataset quality'. One way to address this issue is the GEOSS approach of "fitness-for-purpose", conveying that data quality is in the eye of the beholder. Another possibility is that the system could collect comments from end users about the data. The group identified that there is a relation between data, accessibility and quality assessment and this means that there need to be metrics of the data level. Users should be able to access these metrics so that they can assess which data to use. There need to be standards of the metadata, which are mandatory. If researchers wanted to contribute to data, then there should be recommended formats that have to be accepted and followed. For non-standard datasets, it would be the responsibility of the data owner and the data coordination team to see how best to make that data accessible.

Another issue that was raised is that although the infrastructure might be static, there are different structures that depend on both the software and the format. Given this, participants raised questions about whether there should be a general framework that could accommodate some variety of technological frameworks to support different data and different disciplines. Further, there should be some conditions of use put in place. Quality control is linked to the metadata. Another area that needs to be considered in more depth is Cloud storage, open data infrastructure, persistence and ownership, and changes on the license agreement.

Other points raised include the issue of the necessity of researchers documenting their data practice. This might involve finding some incentives (funders and journals) to change people's behaviour. There should also be a drive to implement simple solutions (software) and services that make it easy to make data open.

Some of the emerging solutions include online tools and software for reproducing data and analysing data (simple workflow service), some of which already exist. These include DOI/URN as a working standard for identification of datasets and for citation. There are also online web services for sustainable storing / working spaces, DRYAD is one such emerging solution. One suggested concept is that a metadata mark and good data visualization tools could be helpful to organise data and make meaning out of it. However, it is thought that developing standards for metadata would take a long time and that some of the solutions might be found in informatics domain.

Other concerns emerged around whether there is enough capacity to make data open over a period of time. The funding model of this is also a challenge because researchers have to pay up front for preservation. Careful thought needs to be paid to access control, some data cannot be made open, which is a particular issue for health data. There needs to be funding to staffing management of access control. Shared services are seen as a possible way of providing more cost-effective services – among a number of institutions or organisations. Another way to cover the cost is to introduce a 'pay per use' policy, thus charging for access to data. Open source software tools are also thought to be sensible in reducing cost and keeping the solutions in the power of the community. Virtual machines are seen as having potential in sustaining such services. For example, virtual machines can be used for security. The data never goes out, but the virtual machine goes in and runs the analysis.

Finally, there is a need to credit original creators for the data, publishers have a role to play in this, and Orchid is a way to identify individuals, further work could be in the area of unique identifiers and data journals.

Concluding remarks:
Rachel Finn, of Trilateral Research & Consulting, introduced the third RECODE work package, which is focusing on legal and ethical issues in Open Research Data Access.

Lorenzo Bigagli, CNR, summarized the conclusions of the workshop, which, among other things, has reinforced the importance of metadata and the need to accommodate heterogeneity of user requirements, making it difficult to identify a unique solution to be adopted. We should build on existing research infrastructure, being flexible applying extensible technological and organizational solutions. Share does not necessarily mean giving for free, so new business models, able to sustain the ORDA approach should be investigated, along with new professional roles stemming from Data Science and Open Data in particular.

He closed with a reminder of the WP2 stand in the EU booth at the GEO-X Exhibition and the RECODE-promoted session ESSI 2.13 at the upcoming EGU General Assembly Meeting in Vienna.

The workshop ended at approximately 17:15.