RDA-WDS Publishing Data Interest Group

Workflows Working Group
Case Statement

Working Group Charter

Objectives

Researchers are increasingly encouraged or required to make their research data available for reuse but might often feel there are insufficient incentives for submitting and publishing data, resulting in low submission rates. Moreover, even when research data are preserved and submitted, it often happens with a bare minimum of metadata which inhibits reuse.

Why is this? There are established and/or emerging workflows for selected disciplines that enable the publishing of data and some provide credit via citation mechanisms. But in most disciplines researchers are simply not aware of such workflows and they may not be applicable without significant modification. Having information about workflows is therefore crucial for researchers—and the people/stakeholders supporting them—to understand the options available to practice open science. Workflows that enable persistence, quality control and access are all crucial to enhance the possibilities for greater discoverability as well as efficient and reliable reuse of research data.

The objectives of this Working Group are to provide an analysis of a representative range of existing and emerging workflows and standards for data publishing, including deposit and citation,

and provide reference models and implementations for application in new workflows. We will report on:

- Investigation and classification of current workflows for publishing data - including a brief gap analysis across disciplines for the identified use cases.
- Identification of a smaller set of reference models covering a range of such workflows to include:
    - when and where QA/QC and data peer-review fit into the publishing process (the broader subject of peer review itself is proposed as a future separate working group)
    - the role of researchers, institutions, data centers, publishers, funders, service providers and the wider community in the data publishing process
    - key barriers for identified use cases
- Selection of key use cases and organizations in which components of a reference model can be applied and promoted to the wider community, working closely with other WGs under the Publishing Data Interest Group.

We will build on the work of major past and current initiatives in which many of the working group members played leading roles. While these initiatives essentially focused on mature examples in particular in the Earth Sciences the work of this group will address a much more comprehensive and multi-disciplinary range of use cases, will classify workflow steps, components, and roles and eventually produce generic workflow models which towards the end of the project will be ready for testing and application in particular in the context of the ICSU World Data System and major science publishers.

Deliverables

- Provide a report summarizing the results of the investigation of current workflows including gap analysis
- A classification of a representative range of workflow models, in each case identifying the varying stakeholders and their different roles and responsibilities, to include where possible the likely associated resource and cost implications (working with relevant proposed RDA-WDS Costs of Publishing Data WG)
- Reference models summarizing key characteristics for each class of workflow
- Implementation of key components of a reference model to an existing use case(s) in order to illustrate the benefits to researchers and organizations of the reference model and the associated implications for the Working Groups on Costs, Publishing Services and Bibliometrics.

Value Proposition

Research communities and their institutions are considering—or in fewer cases are already implementing—workflows on their campuses or using platforms, such as discipline specific or national data centers, to allow their users to share and publish their research data. Many of them have to reinvent the wheel as there is no central resource or knowledge base to guide their efforts. Generic workflows, individual use cases or best practices for publishing data would aid them in establishing appropriate solutions that might include local systems enabling data deposit.

Research data are usually part of a network of scholarly objects, e.g. documentation, lab books or journal articles. It is expected that such clusters of information will continue to evolve and become more complex in the future as users expect to navigate seamlessly within them. They want to discover and access related information without major additional effort. This can only be facilitated by building a detailed understanding of the workflows and publishing outlets available right now. The main challenge will be identifying generic model elements while accounting for discipline specific features. We will cover different steps in the research lifecycle, as needed, e.g. from depositing data in repositories to dedicated data centers, data articles and journals. Identifying the steps in publishing workflows and who is responsible for various tasks can dispel some of the uneasiness for those encountering data publication for the first time as well as offer guidance across emerging and established tools being used in more advanced data sharing communities.

In classifying the current workflows, we will establish general models that allow for the individual imprints from various communities. The result will be reference models and components offering guidance for the wider community, from beginners to more advanced data publishers. This resource will be of use for any stakeholder group involved in publishing data. Repositories are often not aware of journal workflows and vice versa; understanding other parts of this complex endeavor helps each party see its role in the wider context. It is also useful in setting up mechanisms to link data and publications. Librarians have a role here in supporting researchers to find repositories to deposit in that are relevant both to their discipline and their publishing intentions, and offering guidance on the respective journal and repository workflows. The consortium of this working group comprises representatives of all these stakeholder groups to ensure the coverage of the wide range of use cases/best practices and viewpoints already emerging.

One important part of the work in the analysis of workflows will comprise the workflows for the usage of persistent identifiers, in particular Digital Object Identifiers (DOIs), which enable persistent links between digital objects, as well as accurate data citation. Data publication that enables data citation is a key incentive to make data accessible. Furthermore, such persistent identifiers allow an interoperable framework across platforms, publishers, repositories and others.

We plan to build on this in the second phase to test real implementation(s) of generic workflow model components in new scenarios. This offers a mutual benefit for both the provider who can test applicability and promote awareness of tools and for those implementing new workflows and who become able to offer their communities the benefits of publishing data.

Who will benefit

The main beneficiaries of the analysis and subsequent testing provided by this working group are the researchers and the main stakeholders involved in publishing and managing data, as well as in supporting scholarly communication. Better services and strategies for joint workflows will consequently influences the wider research communities. Discoverability and reuse of data will be enhanced, in particular through the unique collaboration between all relevant service groups participating in this group.

Authors will have clear channels for publishing data available to them and, crucially, will be able to derive credit from adhering to best practice in managing and sharing their data. Funders will be able to track the research data they have funded, measure its impact and guard it against repetition. Researchers will be able to work faster and achieve deeper insights outside their immediate subject domain. Librarians and data center experts become an integral part of the

ecosystem, e.g. through their expertise in cataloguing and metadata production and reference models for workflows are templates for ingest, QA, archiving and dissemination. Publishers and other service providers can use reference models for linking data with publications and provide innovative solutions to enhance access to and analysis of the published data. Workflows for data and metadata exchange between the stakeholders who hold it will help policy makers, funders and the public better ensure that the data underpinning published research is being made accessible cost-effectively. Policy makers and the public will be able to navigate the knowledge landscape with increased confidence in its veracity.

Impact

After the first phases, the identified use cases referenced to generic model elements where appropriate, will allow for a unique assessment of data publishing workflows today. This will directly inform and influence the work of all participating stakeholder groups, from repository providers to journal editors. The first steps enable an information exchange beyond the individual stakeholder groups and thus enable the adoption of best practices from other disciplines or joint workflows.

The proposed test implementation(s) of generic workflow model components in new scenarios benefit providers and users as explained previously and enables communities to meet key international and national government mandates to enable and incentivize data sharing for the benefit of all.

Working closely with the associated RDA-WDS Publishing Data Working Groups proposed on Bibliometrics, Costs and Publishing Services, we will identify the role and implementation of emerging metrics and impact/assessment tools in our test workflows and disseminate best practice. This will further the advancement of data aware incentive systems in research.

Engagement with existing work in the area

The WG builds on existing initiatives that have already contributed to a better understanding of workflows across disciplines or institutions for example. In addition, a number of use cases have been identified for the work planned in the forthcoming 1.5 years. Many of the use cases presented below are committed to this working group.

This WG will concentrate on workflows specifically, but linked to the other proposed WGs.
A general overview of relevant initiatives, projects, and platforms will be developed and maintained at the level of the RDA-WDS Publishing Data Interest Group, and may be found in the online survey[1].

In addition, we have been provided with agreements to offer materials by the following stakeholders and will seek further contributions via ICSU-WDS Members and RDA. It is envisioned to expand into a range of disciplines, including further partners in the Humanities and Social Sciences.

[1] Survey of related initiatives, projects platforms: http://goo.gl/0q2f8j

Ongoing initiatives addressing workflows for publishing data:

- The SCOR/IODE/MBLWHOI Library Data Publication Project has developed and executed projects related to two use cases: (1) data held by data centers are packaged and served in formats that can be cited and (2) data related to traditional journal articles are assigned persistent identifiers and stored in institutional repositories. The group has published the "Ocean Data Publication Cookbook"[2]
- The PREPARDE project has been on focused on the implementation of the Geoscience Data Journal (Wiley) workflow from author submission of datasets and papers, through technical and scientific review to publication, including specific data repository workflows at the British Atmospheric Data Centre (BADC) and US National Center for Atmospheric Research (NCAR). The latter incorporated ingestion of data, through data center technical review, to DOI assignment to dataset and bidirectional linking of data papers and datasets.
- KomFor[3] supplies a platform linking research and community based data facilities, libraries and journals. Funded by the German Science Foundation KomFor comprises a long standing consortium of ICSU World Data Centers and the TIB library in Germany collaborating to build sustainable and reliable ways for data publications in line with quality standards in scientific publishing. Part of the previous work of the consortium had been the implementation of DataCite[4].
- The ODIN[5] project studies workflows in two disciplines, the Humanities and Social Science and High Energy Physics, i.e. investigates commonalities and differences. The project has a particular focus on the implementation of persistent identifiers as an enabler in open science. The project is funded under the 7th Framework Program by the EC.

Use cases (some of them already committed to the WG):

- ICPSR (committed): The Inter-university Consortium for Political and Social Research (ICPSR), a repository of social and behavioral science research data established in 1962, has a documented workflow[6] that tracks data from deposit through curation to publication when the data become discoverable with DOIs and accessible on the ICPSR Web site.
- PANGAEA[7] (committed) is a multidisciplinary data center archiving, publishing and distributing geo-referenced data from earth system research. All data sets in PANGAEA are machine readable, citable, fully documented, and can be referenced via DOI. PANGAEA has established workflows for standalone data publications and for data supplementary to a science article. For this purpose cooperation with Elsevier and further science publishers has been built up. A cross-linking service allows to reference supplementary data in PANGAEA directly from the abstract pages of Science Direct or Scopus.
- The MBLWHOI Library (committed) is assigning Digital Object Identifiers (DOIs) to appropriate datasets deposited in the Institutional Repository (IR), Woods Hole Open Access Server (WHOAS)[8]. The Library has also developed a system that automates the ingestion of metadata and datasets from the NSF funded BCO-DMO into WHOAS and

[2] Ocean Data Publication Cookbook: http://www.iode.org/mg64
[3] http://www.komfor.net/
[4] DataCite: www.datacite.org
[5] ODIN - ORCID DataCite Interoperability Network: http://odin-project.eu/
[6] ICPSR workflow http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/ingest/enhance.html
[7] PANGAEA - Data Publisher in the Earth & Environmental Science: www.pangaea.de
[8] WHOAS: http://darchive.mblwhoilibrary.org/

returns a DOI to the data management office. WHOAS has a service with Elsevier to link from ScienceDirect to datasets in the repository and is indexed by Thomson Reuters Data Citation Index.

- The Published Data Library (PDL)[9] is a project of the British Oceanographic Data Centre that provides snapshots of specially chosen datasets that are archived using rigorous version management. The publication process exposes a fixed copy of an object and then manages that copy in such a way that it may be located and referred to over an indefinite period of time.
- INSPIRE physics (candidate): INSPIRE[10] is the digital library serving the global community of High-Energy Physics (HEP).Today, datasets on INSPIRE are treated as independent scholarly records and are assigned a DOI to facilitate their citation. For INSPIRE the next step includes data citations metrics, as well as stronger collaboration with publishers to identify, preserve and display datasets associated with publications.
- Geoscience Data Journal[11], Wiley (committed): Datasets published in this dedicated data journal undergo a peer review process. They are deposited in approved data centers, while being described in short data papers that give details on for example their collection or processing software that was use.
- PENSOFT biodiversity data journal[12] (committed): Community peer-reviewed, open-access, online platform for publishing, dissemination and sharing of biodiversity-related data of any kind. All structural elements of the articles – text, morphological descriptions, occurrences, data tables, etc. – are treated and stored in accordance with the Data Publishing Policies and Guidelines of Pensoft Publishers.[13]
- F1000Research[14] (committed): Life sciences journal - publishes all article types but ensures all articles include underlying data and software where relevant. Uses rapid publication followed by invited but completely transparent post-publication peer review.
- Publishing data in crystallography[15] (candidate, contact: Brian McMahon): Crystallography is a data-rich, software-intensive scientific discipline with a community that has undertaken direct responsibility for publishing its own scientific journals. That community has worked actively to develop information exchange standards allowing readers of structure reports to access directly, and interact with, the scientific content of the articles.
- EBI Genomics + Europe PubMedCentral, UK (candidate, contact: Jo McEntyre): The field of molecular biology has a long tradition of data sharing going back to the open data principles established with the Human Genome Project. Today, this is is a prominent discipline to observe data deluge. At EBI many databases to deposit data are developed and maintained. Researchers identify such datasets in publications by referencing the ID (e.g. accession numbers or DOIs).
- NPG's Scientific Data[16] (candidate - contacts: Ruth Wilson, Susana Sansone): This dedicated data journal published article types, called "Data Descriptors" which are quality assured through peer review process. The Datasets described are deposited in external, community approved repositories. Alongside the narrative Data Descriptor articles ISA-tab

[9] PDL: https://www.bodc.ac.uk/data/published_data_library/
[10] INSPIRE: www.inspirehep.net
[11] http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%292049-6060
[12] http://biodiversitydatajournal.com/
[13] http://www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf
[14] http://f1000research.com/
[15] Publishing data in crystallography: http://dx.doi.org/10.1186/1758-2946-4-19
[16] Nature Scientific Data: http://www.nature.com/scientificdata/

metadata files are produced for each Data Descriptor to provide a machine readable format.

- Earth System Science Data Journal - ESSD[17] (candidate, contacts: Dave Carlson, Hans Pfeiffenberger): This OA journal focuses on the publication of datasets in the Earth Science. The publication process includes an open peer review process. Datasets are submitted to external approved data centers.
- GBIF data publishing[18] (candidate, contact: Vishwas Shavan): GBIF offers a workflow and a toolkit[19] for publishing biodiversity data comprising species occurrences data, species checklists, and corresponding metadata.
- Digital Curation Centre[20] (committed): Centre for advice on research data management to UK universities, including generic advice on support for data selection, deposit and publishing.
- Harvard Dataverse[21] (committed): includes the world's largest collection of social science research data. It supports a full data publishing workflow that can be reviewed and evaluated with a deposit API that integrates with journals to seamlessly deposit data to the research data repository as part of the article publishing workflow.

Work and Adoption Plan

The following tasks will be carried out in close cooperation with the parallel RDA-WDS Working Groups. The work will result in a consecutive set of reports, each of them open to external feedback.

The work is currently envisioned for a timeline of 1.5 years. The success of this working group depends on a close collaboration of all stakeholder groups from data centers to publishers to research community representatives and their institutions etc. We will pursue a wider dissemination and engagement of external initiatives that might emerge over the course of this work. This approach shall ensure an extensive coverage in terms of model components as well as use cases. We then intend to implement one or more model components in suitable use cases which will also require a strong and open engagement with the wider community.

Based on this approach, the work plan is split into four consecutive phases reaching out to mid-2015 on an 18-month timescale. However, given the practical implementation that is planned for the fourth phase in this working group, it is to be expected that work will continue far beyond that time horizon.

PHASE I (end March 2014): Understanding the current state

- Identify a representative range of workflows for publishing data across disciplines: analyze and consolidate them into several broad workflow classes
- Compose questions for use case representatives comparing and contrasting individually selected workflows against a broad workflow class
- Initial discussion of questions at IDCC14 workshop (end Feb 2014), coordination of survey questions with other WGs in the IG Publishing Data

[17] ESSD: http://www.earth-system-science-data.net/
[18] GBIF data publishing: http://www.gbif.org/publishingdata/summary
[19] http://www.gbif.org/ipt
[20] UK Digital Curation Centre http://www.dcc.ac.uk/resources/how-guides/cite-datasets
[21] Harvard Dataverse http://thedata.harvard.edu/

- Present initial synthesis of responses for presentation to RDA Plenary (end March 2014) [**Milestone 1**]
- Identification of groups, workshops and conferences to engage with the wider community and to establish further partnerships, i.e. with disciplinary interest groups.

PHASE II (end of Q3, Sep 2014): Qualifying/Classifying the current state

- Classification and documentation of workflows
    - Define the nature and function of "components" and "workflows" in data publishing
    - identify associated resource and cost implications aligned with proposed Costs WG
    - Identify the varying stakeholders and their different roles and responsibilities: Researchers, Data Centers, Institutions, Libraries, Publishers, Funders & Service Providers
    - Identify where/when/how quality assurance and quality control and data peer-review takes place
    - Identify where/when/how research publishers and journals participate in the data publication process
    - Inclusion of domain coverage, legal and ethical aspects
- Preparation of a draft gap analysis of use cases (e.g. the coverage of components in an individual discipline). [**Milestone 2**]
    - Which components are commonly used, what is missing?
- Wider dissemination and feedback
    - Present draft gap analysis and classification of components: invite additional workflows through RDA community (and discipline specific outlets) and related feedback to refine the models
    - For presentation at RDA Plenary IV (end Sep 2014) and other selected workshops, webinars

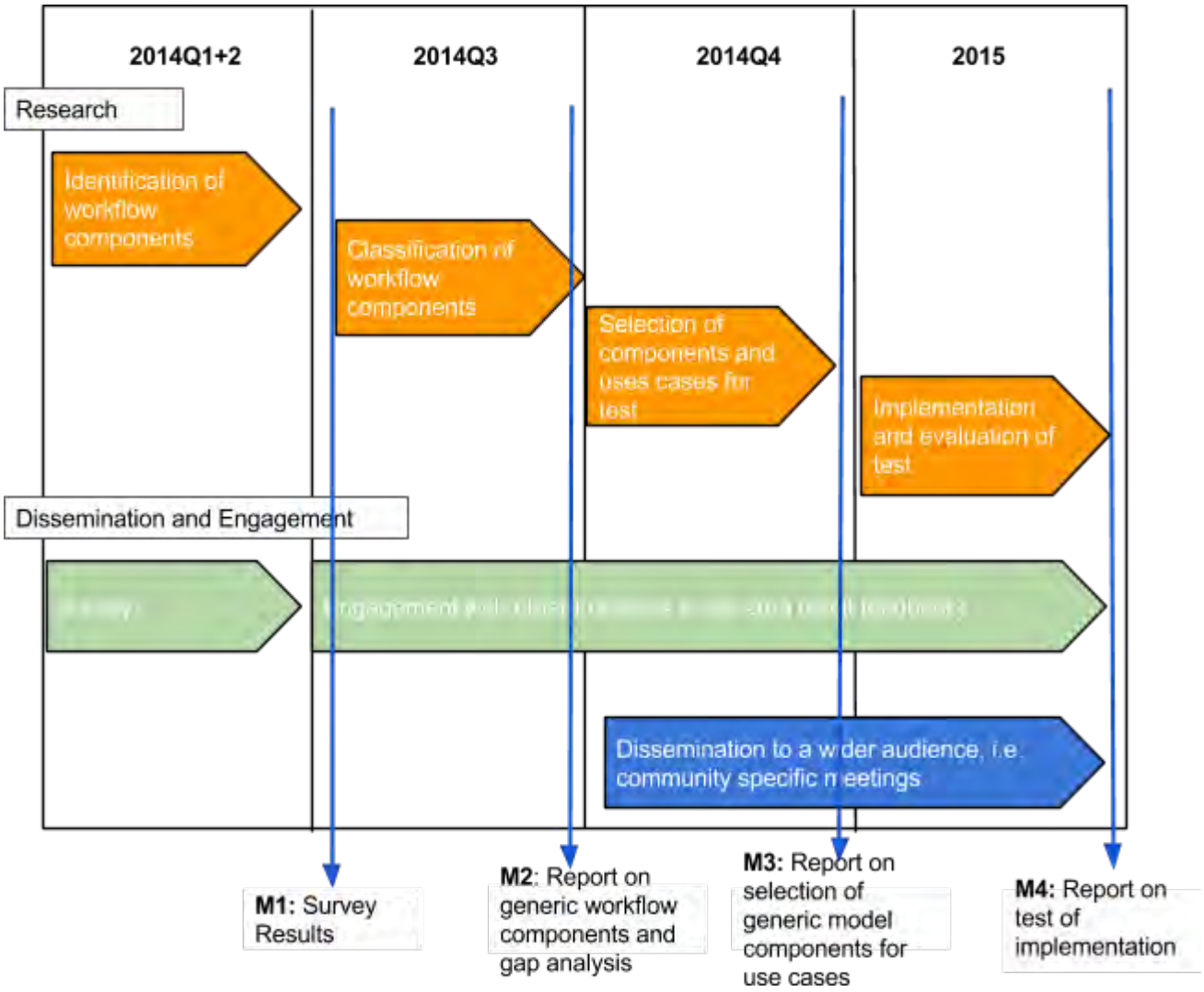
PHASE III (end of Q4, Dec 2014): Final analysis and use case selection for test implementations

- Preparation of final report on workflows, model components and gap analysis which includes feedback from the wider community [**Milestone 3**]
- Based on gap analysis identify and select use case(s) where reference model component(s) may be applied and tested/implemented
    - As part of this: selection of model components in the reference models which may be implemented in a test program for identified use case(s)
- Wide dissemination of results, i.e. also to community or discipline specific groups (possibly also depending on potential test environments)

PHASE IV (end of June 2015): Partnering to fill the gaps - test programme

- Conception of a program with all stakeholders to test one or more generic workflow component(s) in specific use case(s)
- Implement, evaluate and analyze benefits/challenges for use cases (end June 2015)
    - Select a group of potential partners/stakeholders for the implementation of test components in a use case. Questions to consider: What was missing in this use case before? How could the potential partners and model components facilitate data publishing in this institute/discipline/…?

    - Evaluation: how did the test implementation change data publishing in this use case and what were the benefits, including evidence and any available metrics?
- Dissemination within and beyond RDA/WDS



Milestones and intermediate products

- **M1** Survey Results
- **M2** Draft report on reference models, i.e. model components. Special attention is given to a gap analysis of current practices.
- **M3** Final report on components and identification of generic workflow components for test use case(s)
- **M4** Report on test implementation of components in selected use cases: suitability, pros and cons, as well as benefits for identified use case(s)

Project Management

Mode and frequency of operation

The RDA-WDS Working Groups hold face-to-face meetings at RDA Plenary Meetings, teleconference meetings every 6 weeks and builds on regular email communications.

Consensus, conflicts, staying on track and within scope, and moving forward

The RDA-WDS Publishing Data Interest Group will ensure coordination of the Working Groups through regular teleconference meetings of the Chairs of the Working Groups (every 6 weeks), mailing lists and through member involvement in other associated activities, i.e. discipline specific working groups which shall be of special relevance for this topic. The latter will be done through members of the working group as well as further partners who are identified during the work in the forthcoming 1.5 years.

Planned approach to broader community engagement and participation

This working group relies on a strong collaboration of different stakeholder groups. Therefore the engagement within the group and beyond is crucial for its success and will happen through the standard RDA and WDS channels (mailing lists, face-to-face meetings, webinars). In addition, it is envisioned to target community specific conferences, workshops and webinars to engage a broad spectrum of interested parties. This is particularly important when it comes to the discussion of the draft report, as well as the preparation of the test environments. The reports and any other outcome of the working group will be disseminated openly for future reuse/reference.

Membership

- Jonathan Tedds (UK, University of Leicester) [**CO-CHAIR**]
- Suenje Dallmeier-Tiessen (Switzerland, CERN) [**CO-CHAIR**]
- Merce Crosas (US, Harvard University)
- Michael Diepenbroek (PANGAEA)
- Kim Finney (Australia, AADC)
- John Helly (US, UCSD)
- Hylke Koers (The Netherlands, Elsevier)
- Rebecca Lawrence (UK, F1000 Research Ltd.)
- Fiona Murphy (UK, Wiley-Blackwell)
- Amy Nurnberger (Columbia University Libraries)
- Lisa Raymond (US, Library Woods Hole Oceanographic Institution)
- Johanna Schwarz (Germany, Springer)
- Mary Vardigan (US, ICPSR)
- Ruth Wilson (UK, Nature)
- Eva Zanzerkia (US, NSF)
- Angus Whyte (UK, DCC)

References

All of the working groups in the Data Publication Interest Group have a common bibliography[22] in which publications relevant for this particular group are marked correspondingly.

[22] Bibliography: http://goo.gl/wA1G27