New Partnerships for Scientific Data Preservation and Publication Systems

Zhu Zhongming[1], Robert S. Chen, Robert R. Downs,

Christopher Lenhardt, and Xiaoshi Xing[2]

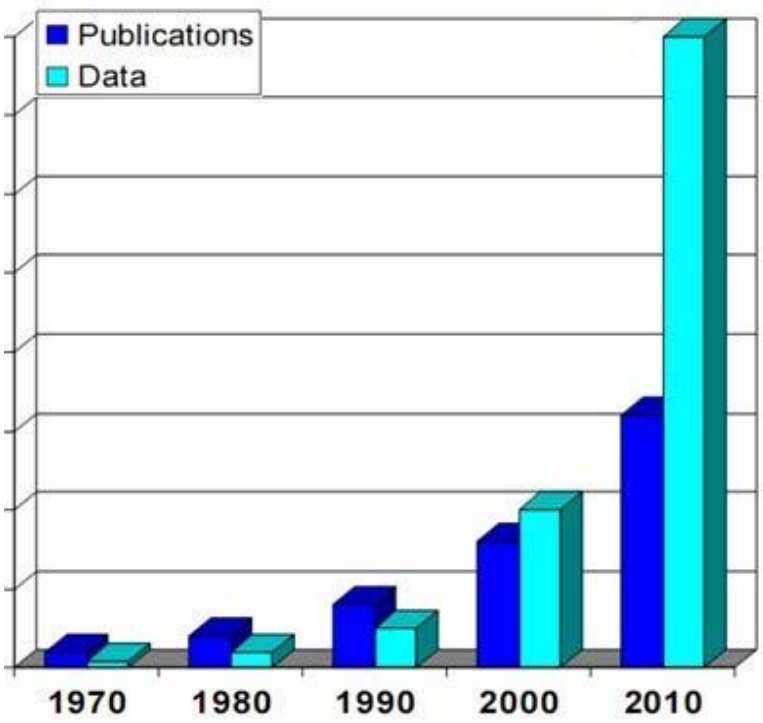
[1]Lanzhou Branch of National Science Library, CAS

[2]Center for International Earth Science Information Network (CIESIN), Columbia University

iPRES 2007, Beijing

Background (I)

- Scientific research is entering a new era of data-driven science



Global Increase in publications in empirical sciences ***(Slightly adapted from [1])***

- In the past years it had been seeing a steadily increase in publications.
- In nowadays, data is becoming increasingly essential part in many disciplines and gaining an exponential growth in volume.
- Moreover, Scientific method is also changing to a manner of "hypothesize, look up answer in data base“[2]

Background (II)

- Scientific data preservation and stewardship is still not optimal and much scientific data remain at risk
  - Whilst loss of published data in some extent are technically solvable, and are the main foci in data preservation community
  - There are still large proportion of never-before-published data left in the dark
    - 80~99% of high-quality scientific data never leaves the laboratory[3]
    - Few publishers support the publication of eData[3]
  - Loss of scientific data appears an inevitable commonplace under current scholarly communication paradigm

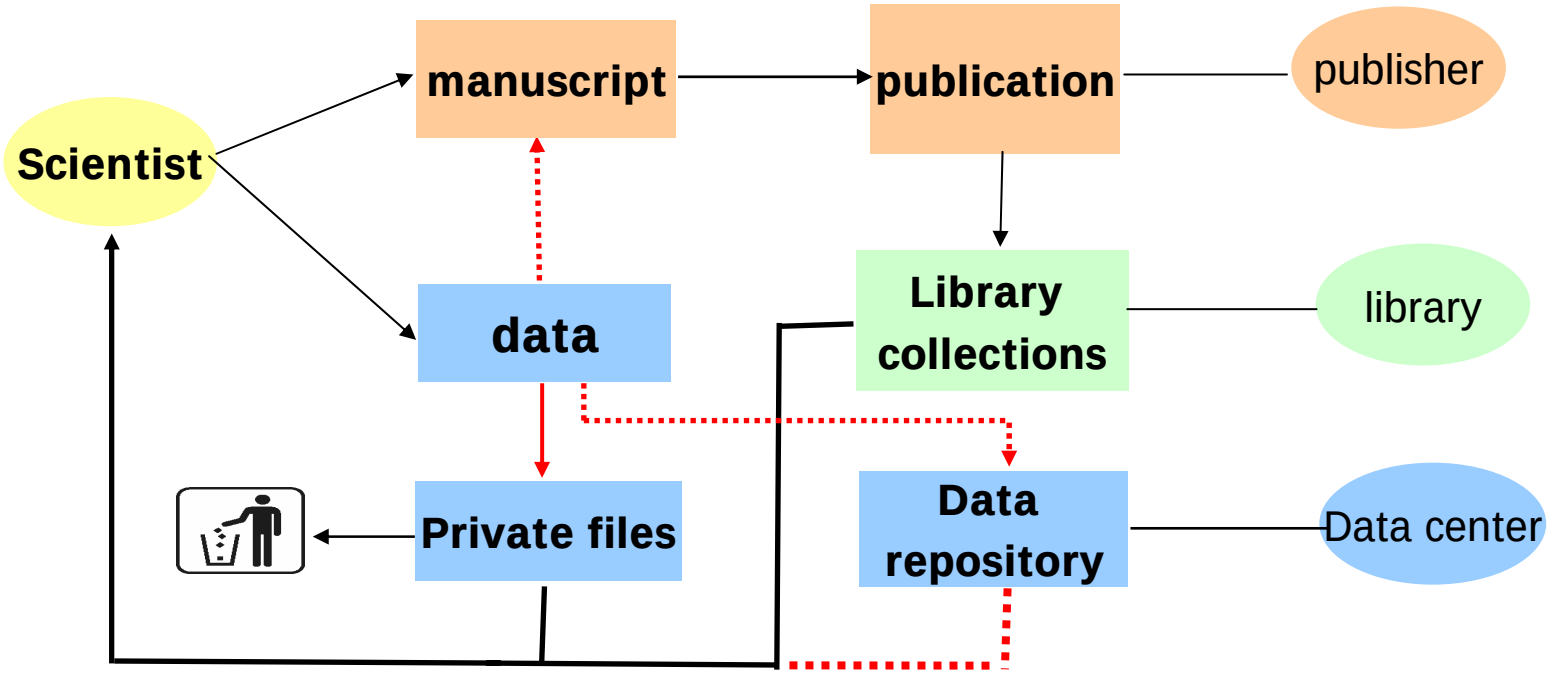
So there is a pressing need …

- To rethink and develop a reliable, robust, and sustainable scholarly communication infra-structure, which
  - is supposed to be a partnership model of promoting integrated publication of research papers and primary data
  - will incentivize scientists to make their valuable data appropriately published, cited and preserved

Let's first look at ……

the current status of scholarly publishing and data dissemination

Data in publication process today[4]

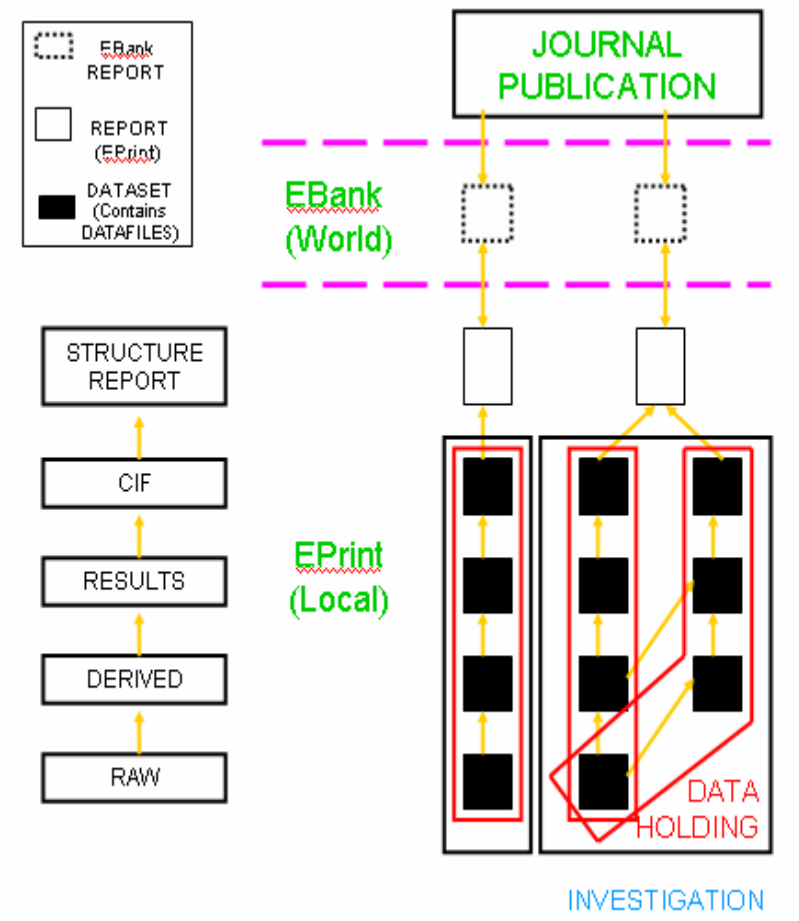


- There is no standard for publishing and citing large volumes of primary data
- Data are separated from associated research papers and are seldom disseminated in an formal publication process.
- Only small portion of secondary or derived data are entering publications
- Some primary data are collected and archived by data centers
- Most are just withhold by individuals and will be trashed or lost by accident.

We are happy to see…..

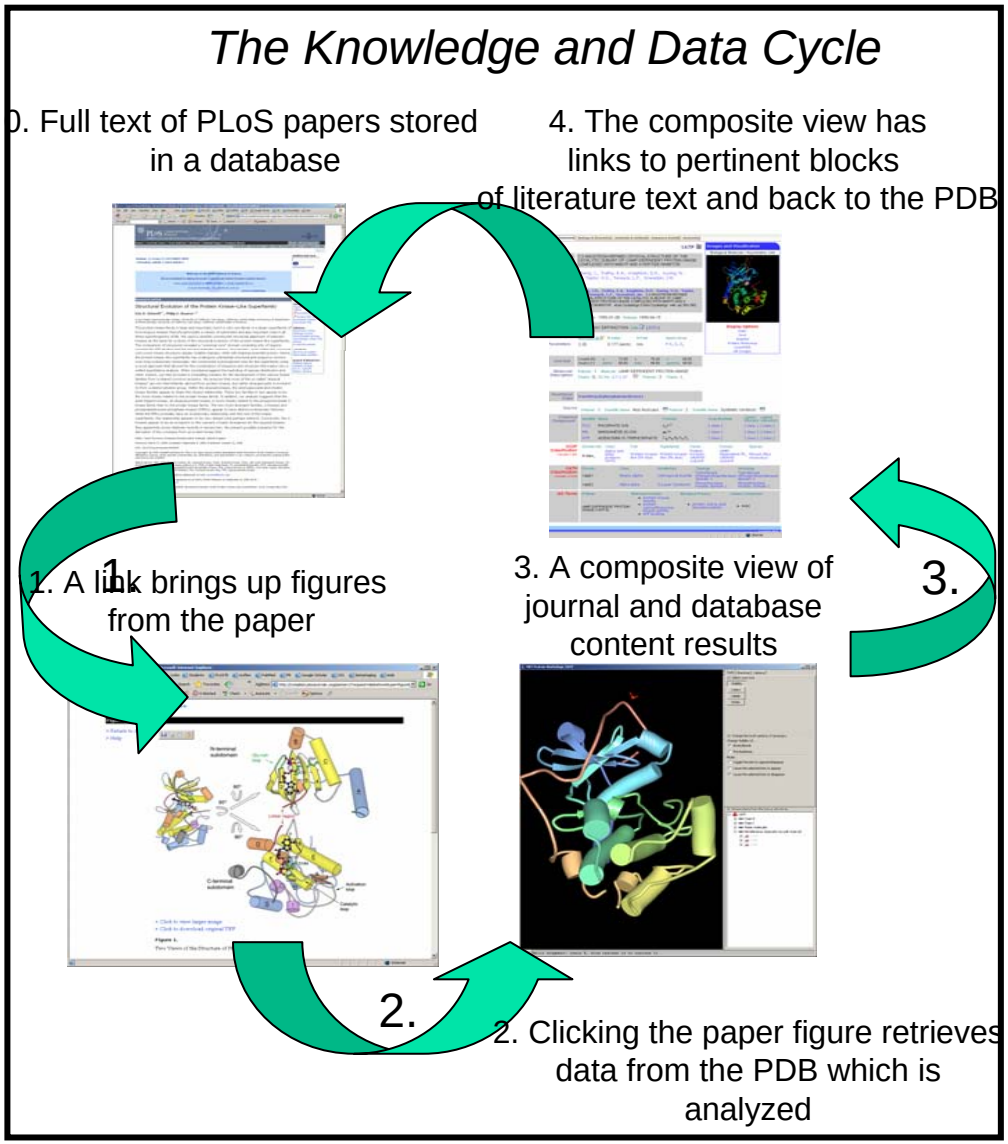
there are initiatives and practices to promote data publishing and linking with publications

eBank UK: Linking Research Data to Publications[5]



- Advocating a 'publication at source' philosophy
- Promote open access to datasets(e-Crystals archive)
- Linking research data to publications and to learning
- Embedding eBank service in research and learning workflows

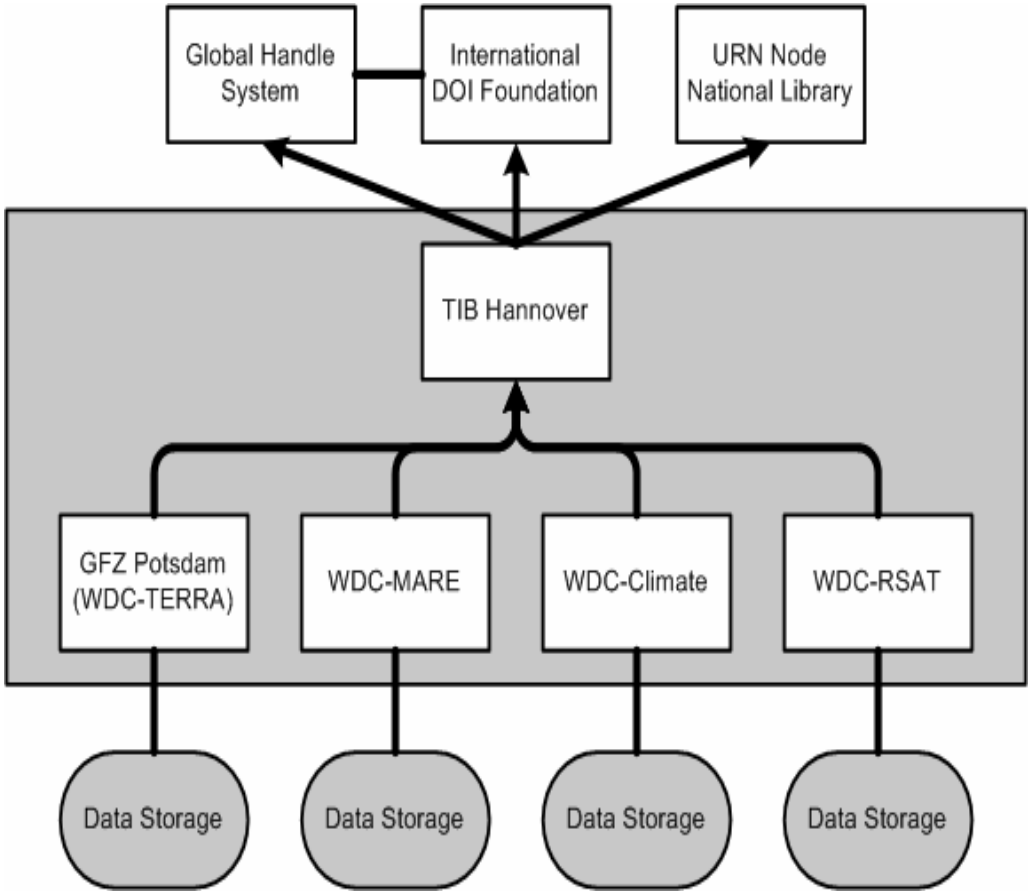
BioLit: Tools for New Modes of Scientific Dissemination[6][7]



A prototyped application:

- Integrates biological literature and biological databases
  - Open access Journals-PLoS journals as literature platform
  - Open archives - Protein Databank(PDB) as linked data repositoy
- Also comprises useful authoring, visualizing tools

STD-DOI - Creating access to scientific Data[8]



- Data publications are processed by publication agents, which are also also responsible for long-term archiving of primary data ("data library").
- Quality control of the primary data and descriptive metadata are set by the author and by the data publishing agency
- Search function for data publications in library catalogues (e.g TIBORDER)
- Access to the primary data with assignment of a persistent identifier and resolver system (DOI resolver)

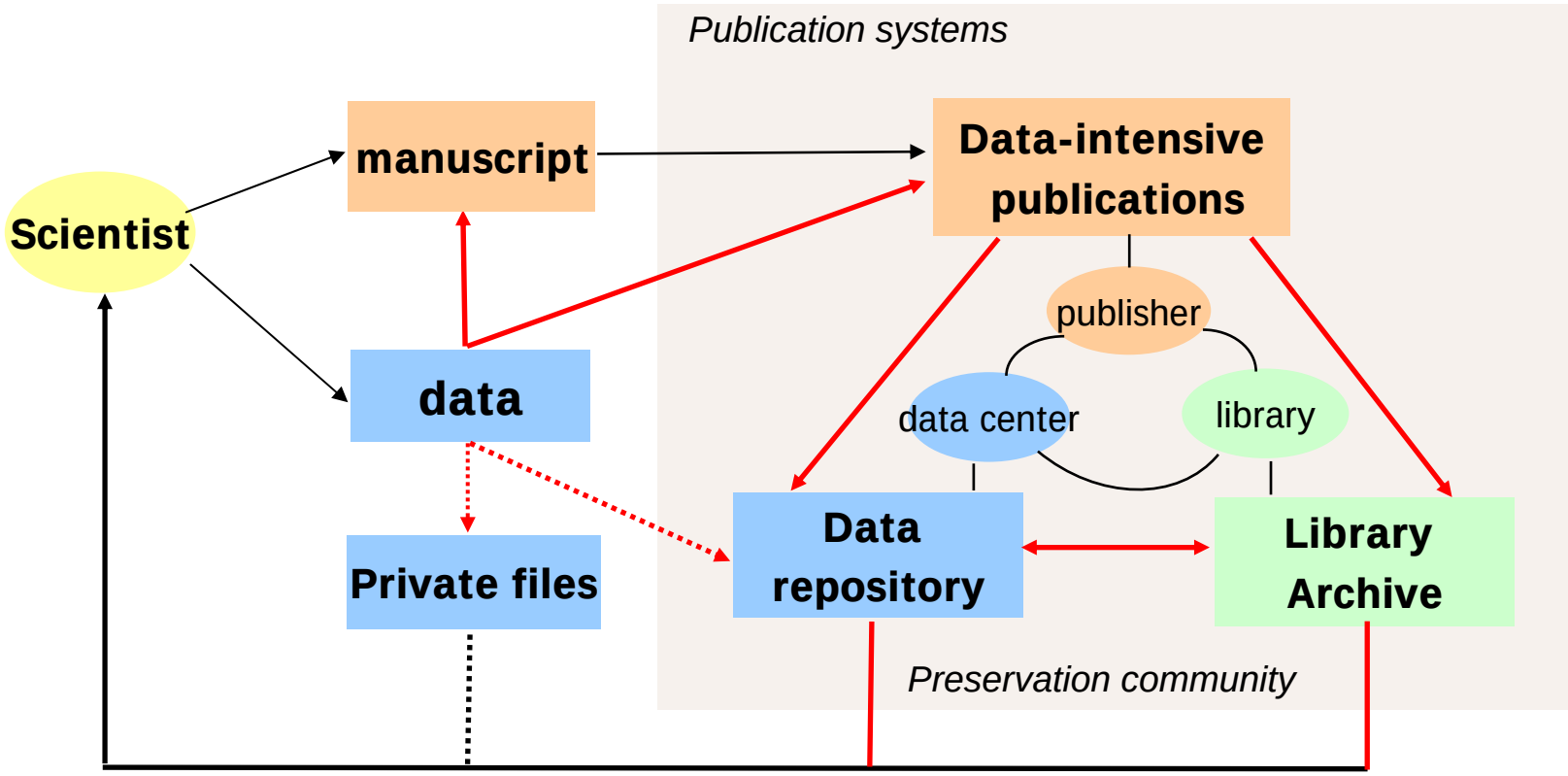
Other similar projects and models

- CLADDIER[9]
  - Funded by JISC, CCLRC library and NCAS/BADC
  - To create an environment and provide mechanics for citing data compatible with citing papers.
  - To provide a mechanism for inter-repositories communication of citation and reference information
- StORe[10]
  - Funded by JISC
  - Source-to-Output Repositories
  - Adding value to the intellectual products of academic research by providing two-way links between source and output repositories
- PANGEA[11] and etc.

Merits and limits of above models

- They have token significant steps in practicing data publishing and shown great potentials in transforming current scholarly publication systems
- Most of them have experimented mutual linking between data and publications
- However they are basically tested within an informal publishing environment
- Further steps need to be taken to bring data publishing into current formal scholarly publishing system, and to build incentive mechanisms to make data publishing and linking a commonplace.

Partnership model supporting integrated publication of research papers and data



- Research data and manuscripts are integratively and simultaneously published
- Research data is supposed to enter formal scholarly communication process
- Valued Research outputs are equally treated and preserved at source.

To make a success story—

- Aligning incentives for data preservation and data publication with those for traditional scientific publications
- Defining roles and responsibilities of stakeholders
- Adapting data management and publication practices

Functions underpinning the success of traditional scholarly publication systems[12][13]

- **Registration:**
  - establishing intellectual priority
- **Certification:**
  - certifying quality/validity of research
- **Awareness:**
  - ensuring accessibility of research
- **Archiving:**
  - preserving research for future use
- ***Rewarding:***
  - evaluating and rewarding performance

Incentivizing mechanisms for data publication and preservation (I)

- Registration:
  - Publisher records the date both the manuscript and associated data was received.
  - Priority claim of data is recognized
- Certification:
  - Introducing a data appraisal and review process
  - Conducted under the auspices of the journal publisher and/or data publication agents.
- Awareness:
  - Via library's harvesting or finding aid to make published journal article and associated data publicly available.

Incentivizing mechanisms for data publication and preservation (II)

- Archiving:
  - libraries and data centers can both provide data archiving service
  - The new model implies, in some extent, the concept of “publication at source” and “preservation at source”. This will definitely mitigate the data loss and make data preservation easier than ever before.

Incentivizing mechanisms for data publication and preservation (III)

- Rewarding:
  - Developing a culture of valuing and giving data the same status as literature
  - Data can be citable, provenance can be tracked and proper credit given to data providers
  - Funding agency, institution and publisher enforce a compulsory policy to make data published and citable.
  - Funding agency and institution reward scientists with an addition of data-citation-impact performance

Stakeholders and responsibilities [15](I)

- Scientists: *creation and use of data*
  - Manage data for life of project.
  - Meet standards for good practice.
  - Comply with funder/institutional data policies and respect IPR of others.
  - Work up data for use by others.
- Publishers: *maintain integrity of the scientific record*
  - Engage stakeholders in development of publication standards.
  - Link to data to support publication standards.
  - Monitor & enforce public standards.

Stakeholders and responsibilities [15] (II)

- Libraries: *Content preservation and awareness service*
  - Collaborate with publisher to provide awareness service concerning journal articles and published research data
  - Provide especially literature-related awareness service, and any other types of scholarly content awareness service
  - Provide preservation service for scholarly content
- Data centers: *curation of and access to data*
  - Manage data for the long-term.
  - Meet standards for good practice.
  - Provide training for deposit.
  - Promote the repository service.
  - Protect rights of data contributors.
  - Carry out data appraisal and review

Stakeholders and responsibilities[15] (III)

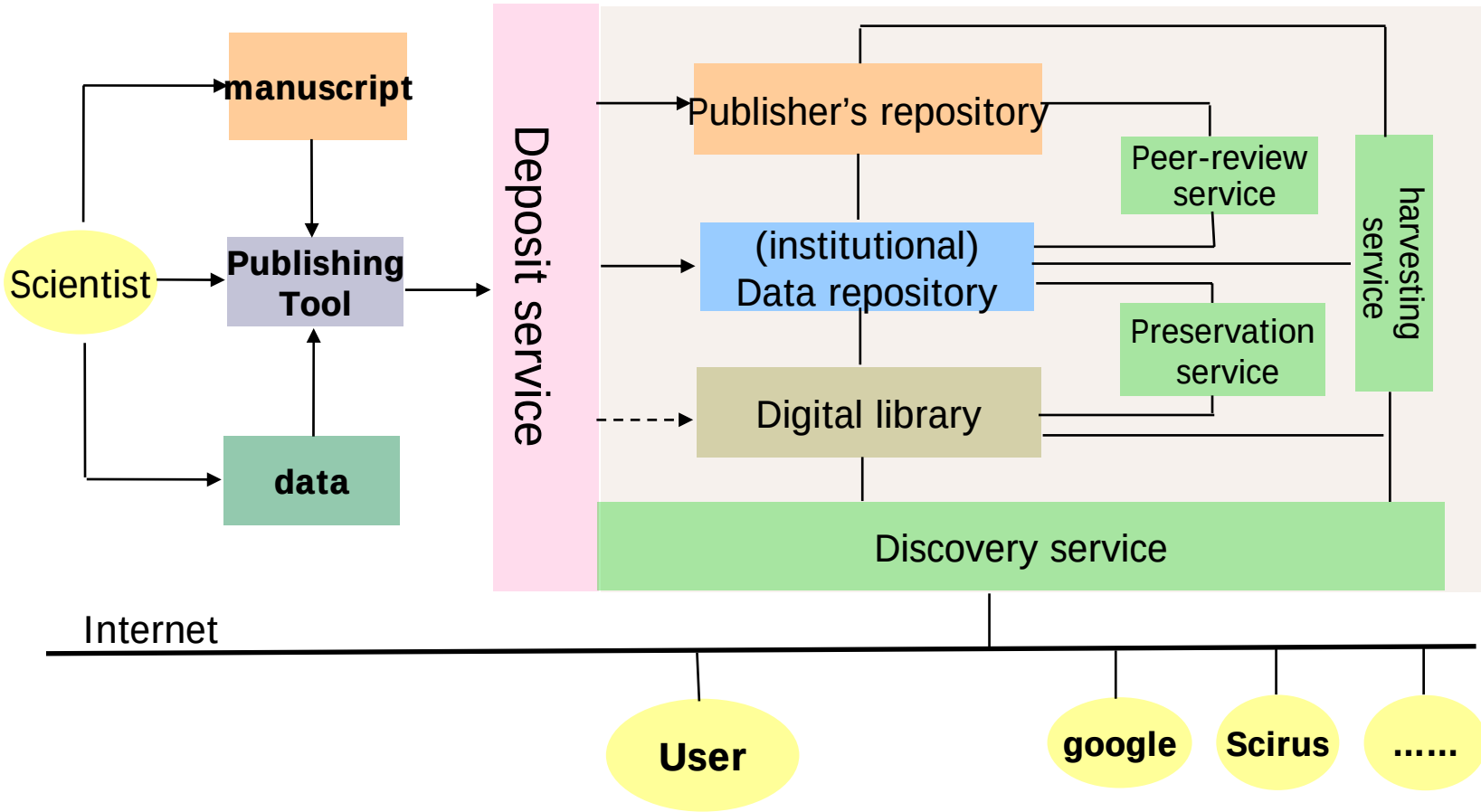
Though below two types of stakeholder are not depicted in the publishing model, but they will play an import role in advancing proposed model to take effect.

- Academic institutions and Funders
  - They both have responsibilities and power in developing policies and practices in relation to how the information outputs from the research are managed and made available.
  - Developing incentive and reward policies and practices to facilitate much more information research outputs are being published and shared.

Adapting data management and publication[1]

- Publish data with associated papers
- Peer review for scientific data
- Make data persistent and citable
- Preservation at source
- Migrate metadata into library catalogues to support common search of scientific data and literature

Proposed technical Framework for publishing data with papers



Framework components (I)

- Manuscript and research data
  - Major research outputs, and inputs to be published and preserved
- Publishing tool
  - Authoring tool particularly for describing and creating metadata for research data to be published with manuscripts
  - Transparent to scientist, that means scientist only need to communicate with publishers via the tool as usual
  - Support of capturing context and provenance information, assigning persistent identifier for datasets

Framework components (II)

- Repositories
  - Publisher's repository
    - Publish papers and auxiliary data
    - Initiate a peer-review process both for reviewing papers and appraising associated data, which may be implemented by data repository
    - Provide brokering service to store and preserve data in data repository
    - Offer harvesting service to disseminate its contents
  - Data repository
    - Should be certified TDR (trusted digital repository)
    - May be institutional-based
    - Publish research data
    - Perform data appraisal and review service on demand
    - Provide data preservation service
    - Offer harvesting service to disseminate its contents

Framework components (III)

- Digital library
  - Aggregate metadata from publisher's repository and data repository
  - Provide discovery service both for papers and data
  - Perform at least literature preservation
  - Probably undertake data preservation task

Framework components (IV)

- Services
  - Deposit service
    - Accept any allowed submission to publisher's and data repository
  - Peer-review service
    - Help to perform peer-reviewing of papers or data appraisal
  - Preservation service
    - Can be called by both data repository and digital library to perform related preservation tasks
    - Provide data provenance monitoring and logging, identification and authentication of digital objects, digital rights management, migration and other functions inside

Framework components (V)

- Harvesting service
  - Enabling service provided by all repositories to share metadata or even full content
- Discovery service
  - Particularly for digital library to be a portal to search and retrieve both papers and data
  - Other two types of repositories own such service interface too.

Data citation mechanism (I)

- Data persistent identifier
  - Uri, urn, ark ... or
  - DOI ?

  Example from STD-DOI[16]:

  doi:10.1594 /WDCC/W_Han_2003_MMB_2

  10.1594

  (Prefix) stands for the TIB as the registration agency.

  WDCC

  stands for the respective research institute.

  W_Han_2003_MMB_2

  is the internal name of the Data

Data citation mechanism (II)

- Universal standard for citing quantitative data[17]
  - Minimal version

Micah Altman; Karin MacDonald; Michael P. McDonald, 2005, "Computer Use in Redistricting", hdl:1902.1/AMXGCNKCLU UNF:3:J0PkMygLPfIyT1E/8xO/EA== http://id.thedata.org/hdl%3A1902.1%2FAMXGCNKCLU

  - Extended version

Sidney Verba. 1998. "U.S. and Russian Social and Political Participation Data," hdl:1902.4/00754 UNF:3:ZNQRI14053UZq389x0Bffg?== NORC [Producer]; data set [Type (DC)] ICPSR [Distributor].

Bibliography

1. Diepenbroek, M., Lautenschlager, M., Paliourras, E., & Grode, H. (2004). World Data center "Earth System Research"-an approach for common data infrastructure in geosciences. Retrieved from http://www.mad.zmaw.de/fileadmin/extern/lectures/CODATAGeneralAssem-Berlin-101104.pdf
2. Lesk M. (2004). Online Data and Scientific Progress: Content in Cyberinfrastructure. http://archiv.twoday.net/stories/337419/
3. Murray-Rust, P. (2005). Open data. http://www.dcc.ac.uk/events/dcc-2006/programme/present-ations/p-murray-rust.ppt
4. Klump , J., & STD-DOI Team. (2007). Semantic linking of data and journal publications in the STD-DOI project. Retrieved from http://www.nesc.ac.uk/talks/712/Klump_GI_2007.ppt
5. Day, M., & Heery, R. (2004). eBank UK: linking scientific data, scholarly communication and learning. Retrieved from http://www.ukoln.ac.uk/projects/ebank-uk/dissemination/jisc-jpm/slides.ppt
6. Bourne, P. (2006). Realizing the Power of On-line Publishing in Copyright and Creativity in International Scholarship. Retrieved from http://www.arl.org/arldocs/resources/pubs/mmproceedings/148/bourne.ppt_files/bourne.ppt.ppt
7. Bourne, P. (2005). Will a Biological Database Be Different from a Biological Journal? PLoS Computational Biology, 1(3). Retrieved from http://compbiol.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pcbi.0010034
8. STD-DOI. http://www.std-doi.de/front_content.php
9. CLADDIER.http://claddier.badc.ac.uk/trac
10. **StORe: Source-to-Output Repositories. http://jiscstore.jot.com/WikiHome**
11. **PANGAEA - Publishing Network for Geoscientific and Environmental Data. http://wiki.pangaea.de/wiki/PANGAEA**
12. Van de Sompel H.(2002). http://public.lanl.gov/herbertv/presentations/OAI-hvds-belgium-brussels-200210.ppt

Bibliography

13. Roosendaal, H., and Geurts, P. 1997. Forces and functions in scientific communication: an analysis of their interplay. *Cooperative Research Information Systems in Physics*, August 31—September 4 1997, Oldenburg, Germany. <http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html>.
14. Lyon,L.(2007).Dealing with Data: Roles, Rights, Responsibilities and Relationships.http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf
15. STD-DOI (2003) Publication and Citation of Scientific Primary Data. Retrieved 2005-09-20 from the World Wide Web: http://www.std-doi.de
16. **Brase,** J. and **Lautenschlager, M. (2004). Pilot Implementation: publication and Citation of Scientific Primary Data. http://www.scor-int.org/ ScientificData-Publishing.ppt**
17. **Altman, M. & King**, G. (2007). **A Proposed Standard for the Scholarly Citation of Quantitative Data. D-Lib Magazine. Volume 13 Number 3/4**

Thank you!