

**BUSINESS CASE – DIGITAL RESEARCH RESOURCE**

**APPROVALS**

**Keith MacGregor**

**Business Sponsor: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_\_\_\_\_

Typed/Printed Name Signature Date

**F P & A Reviewed: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_\_\_\_\_

Typed/Printed Name Signature Date

**TS CFO: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_\_\_\_\_

(≥ $250,000) Typed/Printed Name Signature Date

**TS CEO: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_\_\_\_\_

(≤ $500,000) Typed/Printed Name Signature Date

|  |  |
| --- | --- |
| **Sponsor:** | K. MacGregor |
| **Prepared By:** | Joel Hammond/Bruce Kiesel |
| **Date:** | 7/05/2011 |
| **Revision Date:** |  |

1. **BUSINESS OPPORTUNITY**

This project addresses two key problems in the academic and government market: 1) the difficulty researchers and institutions have in discovering scholarly digital content; and 2) the lack of methods or standards for citing or awarding credit for this work.

Data created for or reused in research studies often sit outside of traditional publications. There are few tools available to aid researchers in discovering these data sets, which could be used for validating research results or re-used in subsequent studies. This project will create a new resource that facilitates the discovery of these data.

Additionally, researchers, university administrators and funders have no straightforward way of attributing credit to digital scholarship. Although citations are widely used to measure the impact of research literature, this measure is not easily or consistently applied to digital scholarship. Building on the Web of Science® (WoS) model and reputation, this project will also focus on capturing bibliographic records and cited references for digital research, and, subsequently, measuring the impact of this growing body of scholarship.

The development of this resource will enable the launch of a new subscription product on the Web of KnowledgeSM (WoK) platform, the Digital Research Resource. Integrated linking of this new resource with other WoK scholarly content, combined with improvements already underway to accomplish full cited reference capture, will improve the overall value proposition of the WoK platform. In the future, SSR’s Business of Science will be able to measure and report on the impact of digital scholarship in InCites, Research in View and custom analytics.

Finally, by exposing a deeper layer of core information in and surrounding the published literature, Thomson Reuters should be well positioned to data mine this research to reveal new hypotheses and claims, materials and methods, and new discoveries.

1. **BACKGROUND and Strategic Rationale**

**The value of digital content to the research workflow**

With nearly ubiquitous access to both computers and the Internet, scholars are forging mini-revolutions in the way they do research and publish. The end of the research process used to culminate with the publication of an article or book. Research data, however, often were discarded or subjected to benign neglect.

Now we see that data produced during the research process are fast becoming first-class objects. Across diverse scholarly communities, many are pointing out that the production and re-use of data should be considered a *scholarly act* much the same way as the primary publications (see Fig. 1). From publishers and funders, we also see a shift in their policies and practices that reflect a change brought on by the use of digital content, including primary research data. A significant push comes from funders who are requiring open access to the data they fund. The funders recognize that there is some urgency since increasingly scholars are seeking access to data that are neither being stored nor curated (see Fig.2).

**Platform Advantage: The Web of Knowledge**

Advances in computational methods have stimulated an increase in the number of scholars within the research community who need to find relevant data sets for analysis, understand how their data are being used, and measure the impact of that use. The worldwide growth in data repositories and mounting pressures from funding agencies to have researchers put their data in them, have added funders and university administrators to a proliferating list of users who need to have a comprehensive view of research data and how it is being used.

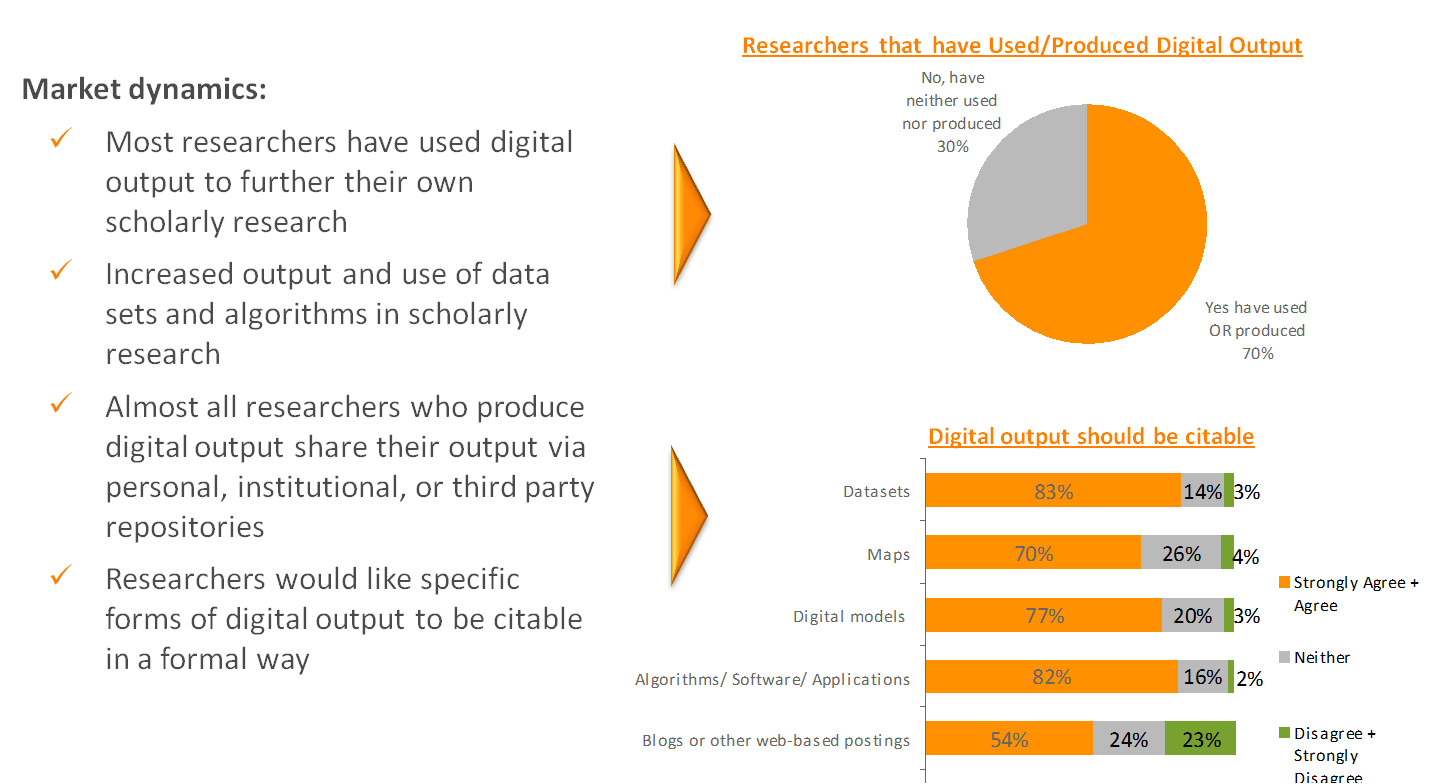


Figure : Most respondents either used or produced digital content and want to have this work cited (Thomson Reuters market survey (2010))

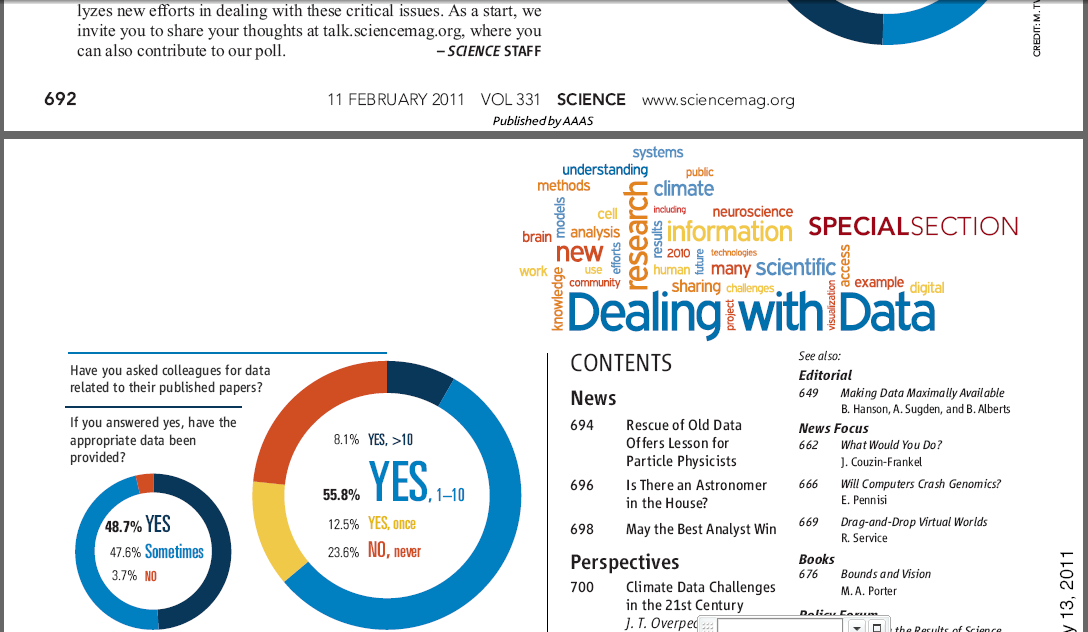


Figure : Increasingly scientists want to find and have access to other researcher's data. (*Science* poll (11 February 2011))

The Digital Research Resource initiative focuses on the deployment of a citation resource that makes digital research discoverable, citable and seamlessly linked to the primary research literature. With the launch of the Digital Research Resource, we will leverage core capabilities of the WoK platform to address identified customer issues regarding discoverability of and credit for scholarly digital content.

The Web of Knowledge platform is uniquely positioned to support the scholarly communications lifecycle and has been developed with the following goals in mind:

* Optimized search and discovery of large and diverse content sets
* Deep connectivity and interoperability across diverse content sets
* Robust citation capabilities that enable search, navigation and attribution of credit
* Standards for tracking and measuring scholarly contributions at many levels, including an individual, institution or work

By analyzing content found in data repositories and feedback from researchers and development partners, we have designed “bibliographic” records that will provide a standard for citing digital research and for collecting citations to that work. In addition, we are collecting the provenance that relates the data set to the research literature, which either produced or reused the data set. The literature provenance will be used to create accurate links between the primary literature in WoS (or other literature collections) and the research data. This connection is essential in that it allows users to contextualize the research data from the perspective of the primary literature or to navigate from the literature to the underlying data.

The Digital Research Resource initially will focus on data sets or data studies deposited in established, curated repositories. The current target list comprises more than 100 repositories, which have a global distribution and provide quality control and editorial policies for managing the data (see APPENDIX B). Also, repositories are in the best position to facilitate long-term preservation and sustainability of research data. Note that the bibliographic model is extensible to other forms of digital content such as maps or videos, which will be added in subsequent iterations.

1. **Project/Initiative Description**

This project has four main objectives:

1. **Identify, select and acquire content** – Identify key repositories in the areas of social sciences, life sciences, and environmental sciences for inclusion in product. Obtain permissions and set up data feeds for each of these repositories. Establish the standards for creating, citing and unifying the “bibliographic” metadata records (e.g., data studies, data sets) for digital research. Develop editorial policies and documentation necessary to support product requirements. Ramp-up the number of active repository feeds, with a target of having ~80 in place by product launch. Identify areas for future coverage expansion, e.g., new media types such as video, maps and images.
2. **Develop required production processes** – Develop, construct and finalize production processes and workflows necessary to support product construction. Develop and finalize Digital Research Resource XML format specifications. Develop process for GEO metadata capture. Architect and establish data feeds to GEO and CMS. Develop CMS process for data ingestion and storage.
3. **Develop the WoK product container** – Rapid prototype to validate the product model. Create and implement required record types. Develop algorithms to facilitate citation unification and metrics. Create links to internal and external collections and repositories. Extend WURS functionality to capture necessary usage data.  Design an effective user interface, based on prototype feedback and usability testing.
4. **Launch product** – Launch a new subscription product on the WoK 5 platform.
5. **AUDIENCE/MARKET**

**Researchers**

The creation of the Digital Research Resource and its integration within the Web of Knowledge addresses the researcher’s need to discover the growing body of non-traditional content. In addition, researchers will be provided with standards for citation, incentives for sharing data, and a mechanism for receiving credit for their work.

**Librarians**

This new resource solves a pain point for librarians who have to help their users identify and find valuable research data associated with the primary research publications. In addition, science librarians themselves use data repositories in their work and need tools that help them find this information. The integration of the Digital Research Resource with WoS and other WoK content improves WoK’s position as a one-stop-shop that addresses core information needs of librarians.

**Funders**

This new resource will enable funders to see the connections among the research data and publications that they have funded. Because WoK citation practices for digital content will provide a standard for citation, funding agencies eventually will be able to measure impact of their funding.

**Research Administrators**

The addition of this new resource to analytic products and WoK will provide essential information to academic administrators and support their need to accurately assess researcher, faculty and institutional performance. The Digital Research Resource will provide a standard for citation, expanding the information administrators can use in measuring research impact and increase their institutions’ competitiveness in winning future grants.

**Bibliometricians**

This new resource will complement existing WoK scholarly literature products and enable bibliometricians to have a more complete picture of scholarly output. This will add an additional dimension to the analytic data we provide and improve our ability to sell high-quality information to this growing and increasingly competitive market worldwide.

1. **COMPETITION**

* **Elsevier**

Elsevier has recognized the growing importance of the underlying research data and has begun facilitating discovery of and connection to these resources in their products and services.

Elsevier’s SciVerse Hub provides one-click access to full text articles (ScienceDirect), abstracts and citations (Scopus), and web content (Scirus). Within ScienceDirect, Elsevier increasingly is exposing the supplemental data, images, video, etc., associated with these full text publications for discovery and reuse.

Elsevier and PANGAEA, a publisher of environmental science data sets, have implemented reciprocal linking, in which research data sets deposited at PANGAEA are linked to the corresponding articles in Elsevier journals on ScienceDirect and vice versa.

Elsevier also recently launched Sciverse Applications, a marketplace and developer network that allows the scientific community to build, find and use applications. Within this marketplace, Elsevier has created an application that, for articles having supplementary data in PANGAEA, displays a Google map indicating the geographical coverage of each dataset

Another ScienceVerse Application, the US Government Dataset Search application, automatically displays the results of a search of US government datasets as the user runs a search on SciVerse Hub.

* **Individual Data Repositories**

In addition to providing search and discovery solutions themselves, many of the publicly available data repositories make their data available for indexing and discovery by internet search engines such as Google and Bing. Increasingly data repositories are providing connectivity to other web resources, including the primary research literature. Among the leaders in this area is the National Library of Medicine’s National Center for Biotechnology Information (NCBI). NCBI already offers tight integration among the various databases for which they have responsibility, e.g., Genbank, PubMed, PubMed Central, as well as links to other web resources. Domain-specific repositories exist worldwide and are federating connections among themselves. What these organizations lack, however, is the ability to get a comprehensive view to assess the impact of this work.

1. **SCOPE OF PROJECT/DETAIL DESCRIPTION**

**Assumptions:**

* 2 million records available at product launch; .5 million records/yr ongoing
* All data for GEO and CMS processing will be in XML
* Record Stats:
  + Average # of authors: 6/record
  + Average # of affiliations 2/record
  + Average # of cited references: 3/record

**Key Requirements for Digital Research Resource (DRR):**

| **Requirement** | **UK Editorial Development** | **GEO** | **CMS** | **WoK: UI, Middleware, SES,** |
| --- | --- | --- | --- | --- |
| Development of XML schema | X |  |  |  |
| Acquisition and normalization of digital content metadata from web repositories | X |  |  |  |
| Process cited reference data to match standard WoS citation processing |  | X |  |  |
| Provision for regular receipt and updating of normalized, schema-compliant XML records | X | X | X |  |
| Construction of DRR database in CMS |  |  | X |  |
| Construction of a new DRR container on WoK 5.x |  |  |  | X |
| General search, e.g., search by keyword |  |  |  | X |
| View, print, export, etc. “bibliographic” full records of all record types, i.e., repositories, data studies, datasets and microcitations |  |  |  | X |
| Dedicated Marked List with output options |  |  |  | X |
| Query summaries with search refine capabilities |  |  |  | X |
| Results analysis |  |  |  | X |
| Alerting support |  |  |  | X |
| Navigate between the records, based on the relationships present in the data |  |  |  | X |
| Link to WoK articles that cite/ acknowledge datasets |  |  |  | X |
| View citing articles in WoK from associated datasets |  |  |  | X |
| Link to external Web pages for repositories |  |  |  | X |
| Link to download datasets |  |  |  | X |
| WURS tracking: links, record exports |  |  |  | X |
| Cited reference search |  |  |  | X |
| Display citation counts for DRR records in the WoK Citation Scorecard |  |  |  | X |

1. **MILESTONES** 
   * + Data acquisition
     + Editorial policy development
     + Go-to-Market strategy
     + Production Workflow and infrastructure development
     + WoK interface development
     + Product Launch

**APPENDIX B: REPOSTIORY MAP**

The map below shows the geographic distribution of more than 100 data repositories identified for inclusion in the Digital Research Resource. As a proof-of-concept, data feeds have been established for 19 repositories (Green; listed below). Another 20 repositories are under active discussions (Yellow), and approximately 65 repositories remain to be contacted (Red).

Geographic Distribution of identified data repositories



Data repositories (19) used to evaluate content for the Digital Research Resource

| **REPOSITORY** | **LOCATION** | **CONTENT** | **SUBJECT AREA** |
| --- | --- | --- | --- |
| Australian Social Science Data Archive | Australia | 2,118 studies | Social Science |
| CaArray | USA | 169 records | Life Science |
| CanGEM | Finland | 26 Studies | Life Science |
| CPLA | China | 3,311protein records | Life Science |
| DataCite | Germany | 1 million data sets | Social Science |
| DataWeb | USA | 405 studies | Social Science |
| Dryad | USA | 7,500 datasets | Life Science |
| Finnish Social Science Data Archive | Finland | 739 studies | Social Science |
| Gene Expression Omnibus | USA | 22,261 studies | Life Science |
| Greegenes | USA | 752,661 DNA records | Life Science |
| Interuniversity Consortium for Political and Social Research (ICPSR) | USA | 7,700 studies, 62,000 datasets | Social Science |
| Institute for Quantitative Social Science (Harvard University) | USA | 6,689 studies | Social Science |
| Microkit | China | 1,489 protein records | Life Science |
| Odum Center | USA | 3,194 studies | Social Science |
| PANGAEA | Germany | 2,719 studies, 62,900 datasets | Environmental Science |
| Recode | USA | 1,292 records | Life Science |
| Roper Center | USA | 12,064 studies | Social Science |
| UK Data Archive | UK | 4,964 studies | Social Science |
| UniProt (Universal Protein Resource) | Europe | 2,587 records | Life Science |