Curating Published Data

Mark Cyzyk
David Reynolds

The Sheridan Libraries
Johns Hopkins University

Science, Verifiability, Reproducibility, Data

Two of the fundamental tenets of the scientific enterprise are verifiability and reproducibility of results

In many cases, verifiability and reproducibility cannot be accomplished without access to the original data

Hence the need to collect, curate, and provide access to the original data

Johns Hopkins' “DataPub” Project

Create proof-of-concept system:

A system for enabling submission of associated datasets to a publication system

A system for harvesting and preserving published articles and associated datasets from a publication system

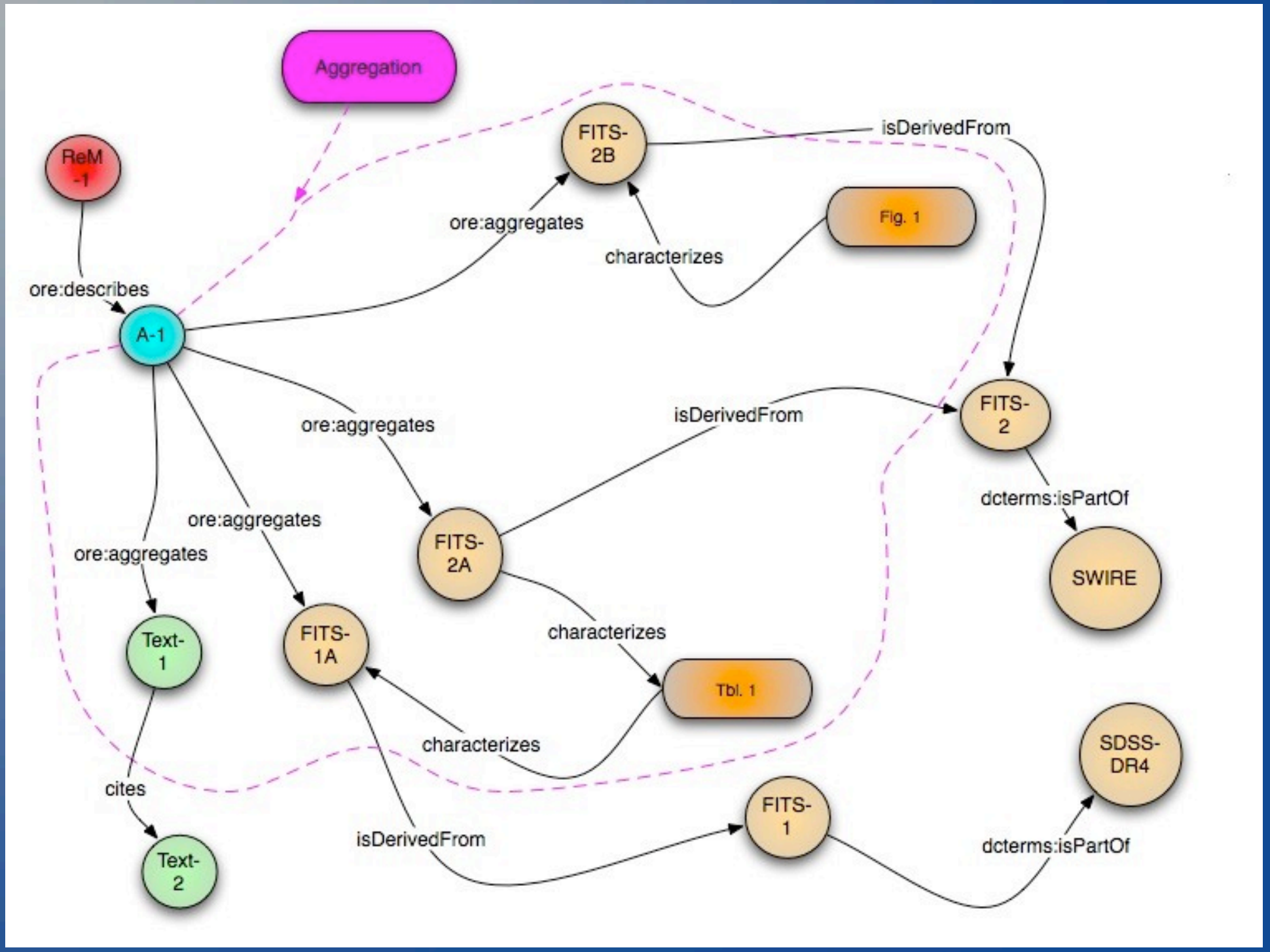
A system for capturing and preserving the relationships between the published article and its associated datasets

Linkage: Article and Data

The need to capture relationships: “This thing is linked to that thing in this way.”

Object Reuse and Exchange (ORE) Aggregation

Resource Map (ReM) – RDF/XML serialization


Aggregation
ReM
-1
FITS-
2B
isDerivedFrom
ore:aggregates
Fig. 1
characterizes
ore:describes
A-1
ore:aggregates
isDerivedFrom
FITS-
2
dcterms:isPartOf
ore:aggregates
ore:aggregates
FITS-
2A
SWIRE
Text-
1
FITS-
1A
characterizes
Tbl. 1
characterizes
SDSS-
DR4
cites
FITS-
1
isDerivedFrom
dcterms:isPartOf
Text-
2

Example Publisher ReM

PublisherReM.xml

http://jhepp.library.jhu.edu/ojs23/index.php/42/article/downloadSuppFile/43/67

Flow of Content

Author submission Web app – assemble ReM

SWORD submission to Publication System

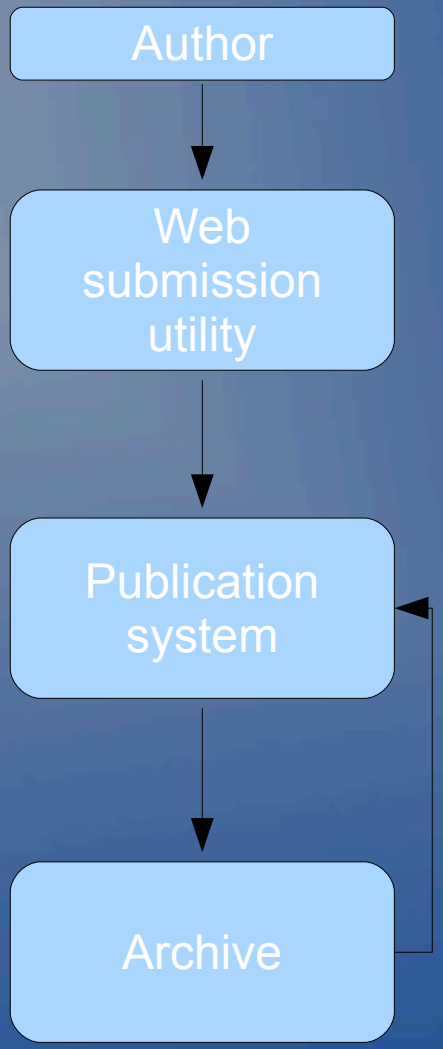
Publication System internals

Archive-side harvesting of content

Archive-side push of transformed ReM

Flow of Content

Author submits article and files to Web utility



Web utility creates initial ReM, packages everything up in Zip file, and makes deposit via SWORD

Publication system accepts SWORD deposit, opens Zip, inserts submission into regular internal workflow

Archive harvests published articles, supplementary files, and Publisher ReM, stores datasets locally, then feeds back a processed Archive ReM with links pointing to objects in Archive

The Archive: Home of Data

Provides permanent home for datasets

Datasets can be interlinked

Datasets can be cited

Citation analysis of datasets made possible

Discovery tools can be built

Thank you

Johns Hopkins thanks both the Institute for Museum and Library Services (IMLS) and Microsoft Research for providing support and funding for this project.

More Thanks

More thanks go to Bob Hanisch and Ani Thakar

The DataPub Team:
Sayeed Choudhury
Mark Cyzyk
Tim Dilauro
Elliot Metsger
Mark Patton
David Reynolds
Cynthia York