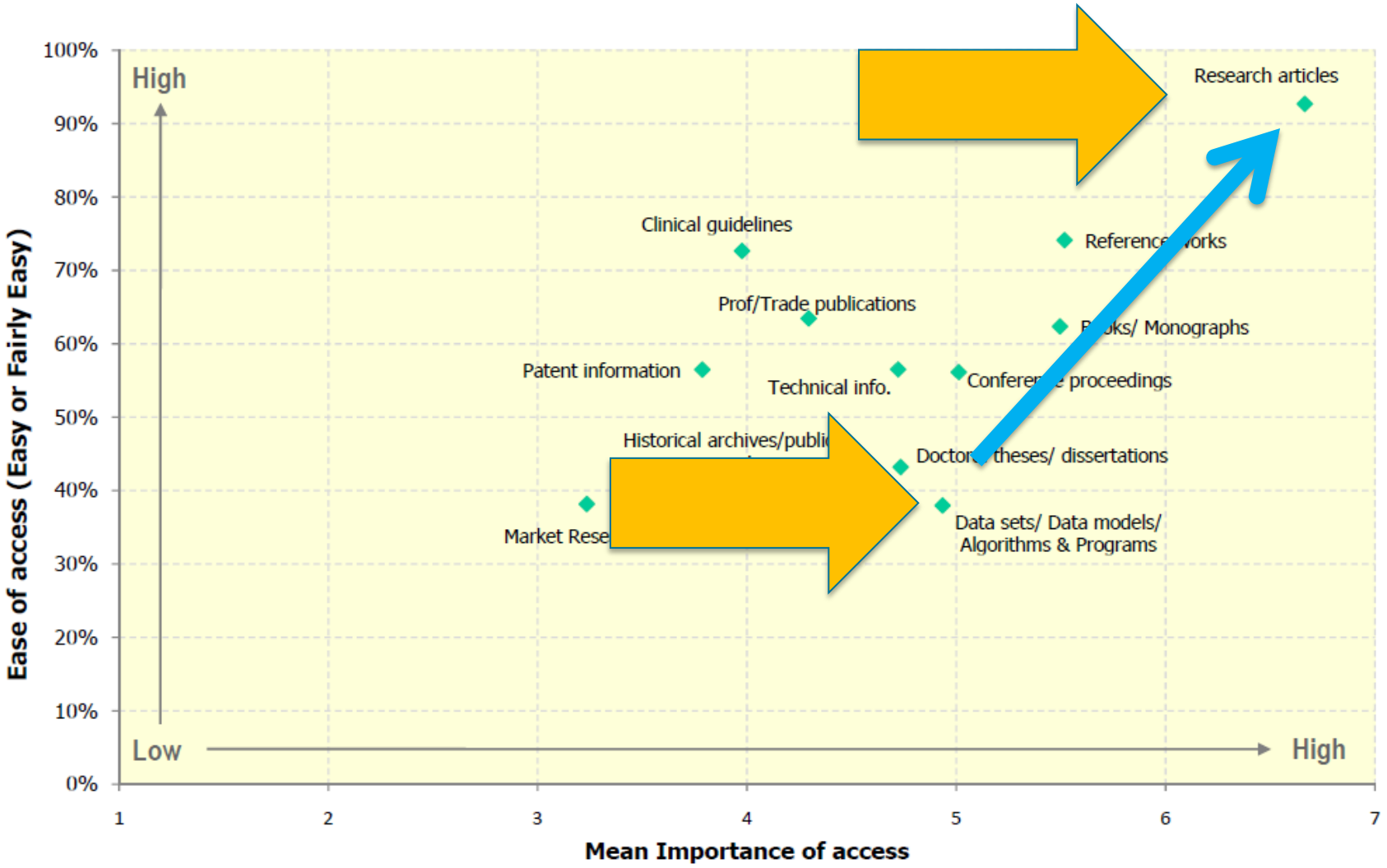
An Open Linked Data Repository for Article-Data Associations (OLDRADA)

A demonstrator in the context of the National Data Service

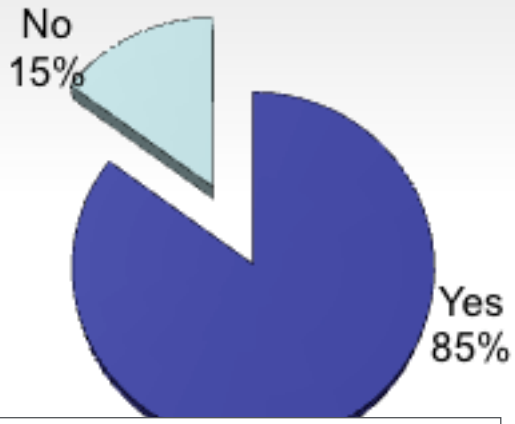
IJsbrand Jan Aalbersberg
SVP Journal and Data Solutions, Elsevier
NDS Rockville, October 23, 2014

Over time, data sets grew in importance and availability, however they were difficult to find

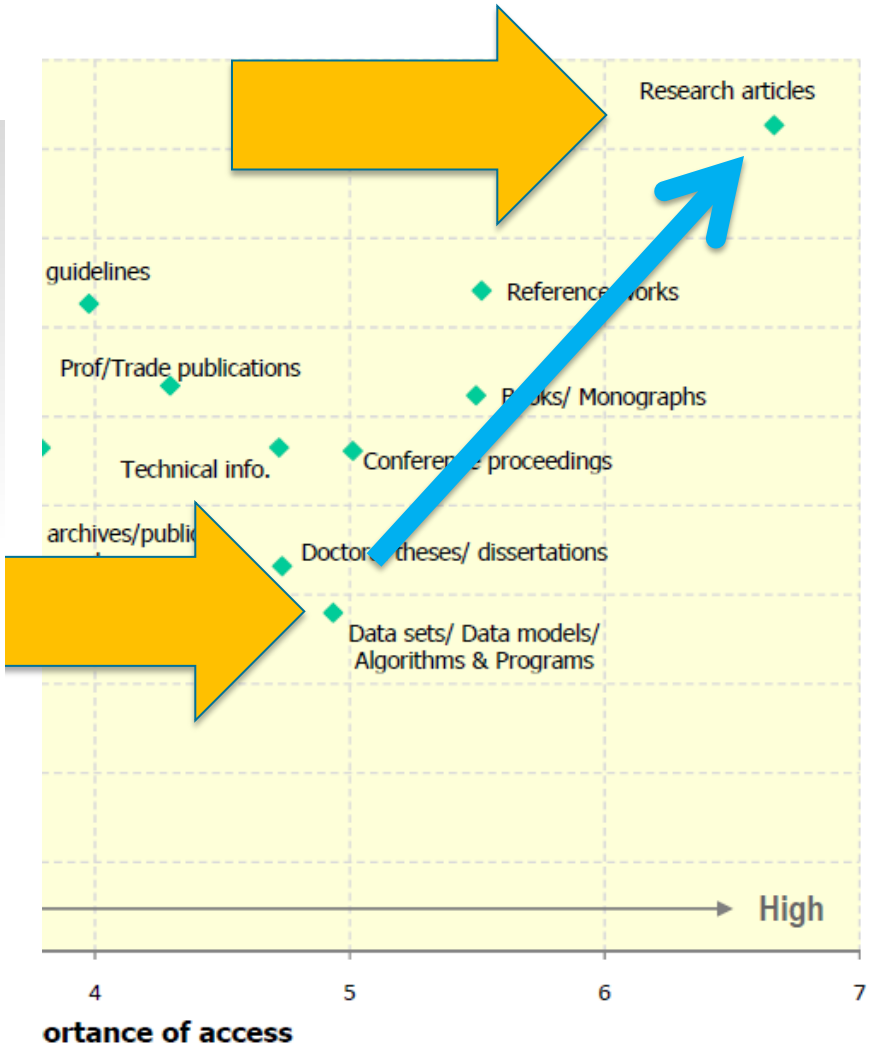


Over time, data sets grew in importance and availability, however they were difficult to find

Do you think it is useful to link underlying research data with formal literature?



Researcher survey, 1202 respondents (PARSE.insight 2010)



Can we use journal article's ease of access to better discover data sets in data repositories?

ELSEVIER

- Formal published articles and their associated research data is often disconnected.
- In some cases one-way links exist:
  - Articles have links to (ID's of) datasets or
  - Datasets have links to (ID's of) articles
- However, this info is usually not shared and often these links remain uni-directional.
- Need for a central deposit and look-up that connects articles with associated research data.
- Topic is also being investigated by RDA/WDS "*Data Publication Services Working Group*" and *Rmap* project (Portico / IEEE / J Hopkins).

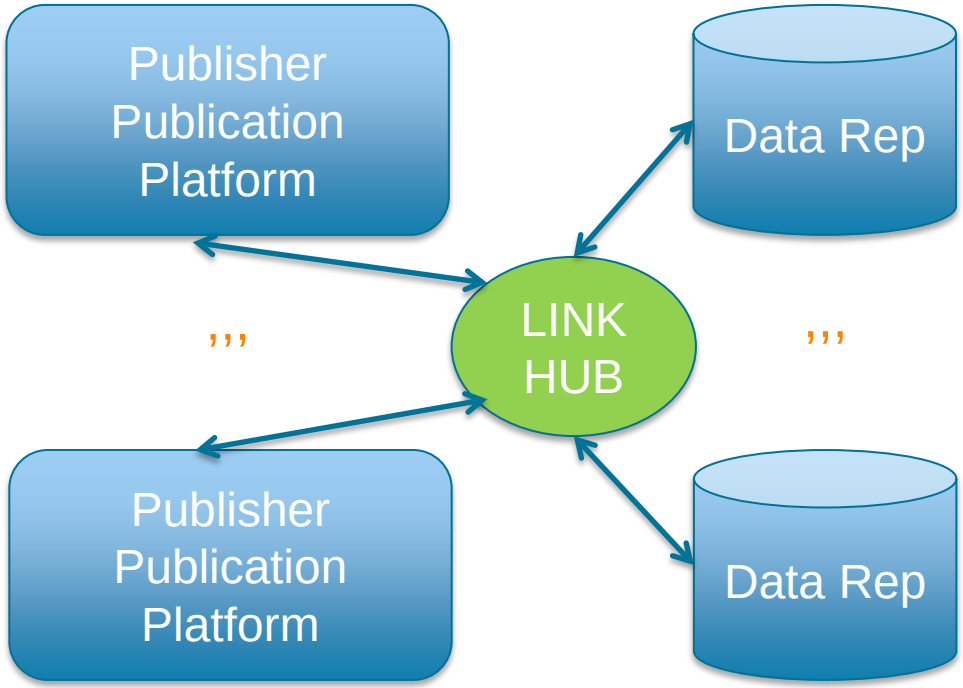
National Data Service (NDS) is: http://www.nationaldataservice.org/

- A vision of how researchers across all disciplines find, reuse, and publish data.
- An international federation of data providers, data aggregators, publishers, community-specific federations, and cyber-infrastructure providers.
- It builds on data archiving and sharing efforts under way and links them together.

- It is a community-driven effort and called for short-term creation of demonstrators to contribute to development of NDS vision.

An Open Linked Data Repository for Article-Data Associations (OLDRADA)

- OLDRADA is an (intermediate) demonstrator to bridge the gap between the short-term NDS call and the medium-term (and final?) RDA solution.



<u>Keywords</u>:
short-term pilot,
open linked data,
partners,
learn,
and share

Partners from data repositories and publishers work an ambitious timeline

ELSEVIER

Article Publishers:
- APS
- NPG
- PLoS
- Elsevier

Data repositories:
- Globus
- MGI
- Figshare
- IEDA
- DataCite
- PANGAEA

Others:
- IEEE (observer)
- STM

Deliverable	Original Timeframe
Scope and planning	July 15 – August 1
User stories collection and prioritization	August 1 – August 31
Data model definition	August 15 – September 15
Access API definition	September 1 – September 30
Development sizing	September 15 – September 30
Development GO/NOGO decision based on cost estimates	
Development	October 1 – November 30
Partner testing	December 1 – December 31
Operation	January 1 – December 31 (2015)
Share lessons learnt	January 1 – December 31 (2015)

as a publisher,	I want to query the system with my deposit ID as input.
as a data center	I want to deposit a batch of article-ID / dataset-ID associations.
as a data center	I want to modify or remove an association previously deposited by me.
as a data center	I want my dataset to be identified by a URL or by base-URL plus ID.
as a project partner,	For one of my ID's I want to know how often it has been used as input paramater.
as a project partner,	Datasets need to have a type associated (some limited provenance and availability).
as a project partner,	In any results list, I want the results to be completely deduped.
as service operator,	I want to variably manage the incoming requests.
as a data center	I want to search using basic information like titles, autors and funders.
as a publisher,	I want an embargo period on data-article associations.
as a publisher,	I want all my datasets to be fully discoverable (e.g., from Google)

Questions (required decisions) that arise from ~25 resulting (deduped) needs

ELSEVIER

- Linking service of full discovery service?
  - E.g. “I want Google to index OLDRADA”
- Amount metadata that needs to be stored?
  - E.g. “I want to search object titles etc”
- Who is allowed to submit links?
- Who guarantees the quality of links?
- How easy is it to update the links?
  - Ingestion process – crawl, upload
  - Changes and versioning of datasets
- How limited is the scope of this going to be?