Towards a Model for FAIR Data Information Infrastructures

Caspar Terheggen, Ed Simons, Radboud University

Radboud University

IN·DEI·NOMINE·FELICITER

SCOPE OF THE PRESENTATION

- About Research Data *Information* Infrastructures, dealing with aspects captured by the term "FAIR":
  - *Findability*
  - *Accessibility*
  - *Interoperability*
  - *Reusability*

  of datasets
  Where I would like to add the term "*Interpretability*" for the "I" of FAIR.

- So it is NOT about data storage infrastructures.

- The aspect of "FAIR" has to do with (optimal supply and availability) of METADATA about datasets.

POINTS OF DEPARTURE

- *Silo-ed registration of metadata on datasets*, meaning in a separate system, only dealing with dataset metadata out of their broader context *is not optimal*.

- A model concerning *FAIR data infrastructures should include the local (institutional), national, international and discipline level.*

- An optimal FAIR data infrastructure consists of two complementary parts:
  - *A technological or “systems” part.*
  - *A service and support part.*

- *FAIR data information infrastructures and services* (built upon these infrastructures), *can in principle be developed and managed by different organisations /entities than the ones that deal with storage infrastructures.*

FAIR Model – Technological Part: THE ROLE OF CRIS's

- Silo-ed registration of metadata on datasets, meaning in a separate system, only dealing with dataset metadata out of their broader context is not optimal and should be avoided.

- Why is this so?
  - *Combining dataset metadata with additional metadata* on (the related) publications the researchers involved, the project the dataset resulted from, the institute(s), the cooperations, the funders, etc… *provides a much richer information source for "FAIR-ness" of datasets* and – not a minor point – for finding interesting related datasets.

- Given this CRIS's, and registering metadata on datasets in the *CRIS, can (and in my view should) take a central position in FAIR data infrastructures.*

THE ROLE OF CRIS's: a bit of history

- CRIS's started (1990's) as administrative, reporting systems to the government, and *were mostly conceived by researchers as "an administrative overhead and nuisance".*

- In the course of the years 2000: CRIS's became instruments for research policy and management and also as a basic resource for institution's OA Repositories.

- The last 5-10 years: CRIS's being used as instruments for on line profiling of research(ers) and research institutes. As a result: CRIS's more and more accepted by the researchers.

To summarize*: CRIS's have evolved* from systems for administrators and managers *to (also) interesting tools for the researchers themselves*.

THE ROLE OF CRIS's: a bit of history

- The researcher's "traditional" view on CRIS's.

Research activities

"My World"

Research information activities (CRIS)

NIH Budget History
NIH budget mechanism detail FY 2000 - 2009
View
NIH budget
View
Total NIH bu
View
Total NIH bu
View
Research Grants
Small Business Research (SBIR / STTR)
Success Rates and Funding Rates
The NIH-Funded Research Workforce

"Their (administration) World"

THE ROLE OF CRIS's: a bit of history

- Is changing....

Research activities

Research information activities (CRIS)

NIH Budget History

NIH budget mechanism detail FY 2000 - 2009 View

NIH budget View

Total NIH bu View

Total NIH bu View

Research Grants

Small Business Research (SBIR / STTR)

Success Rates and Funding Rates

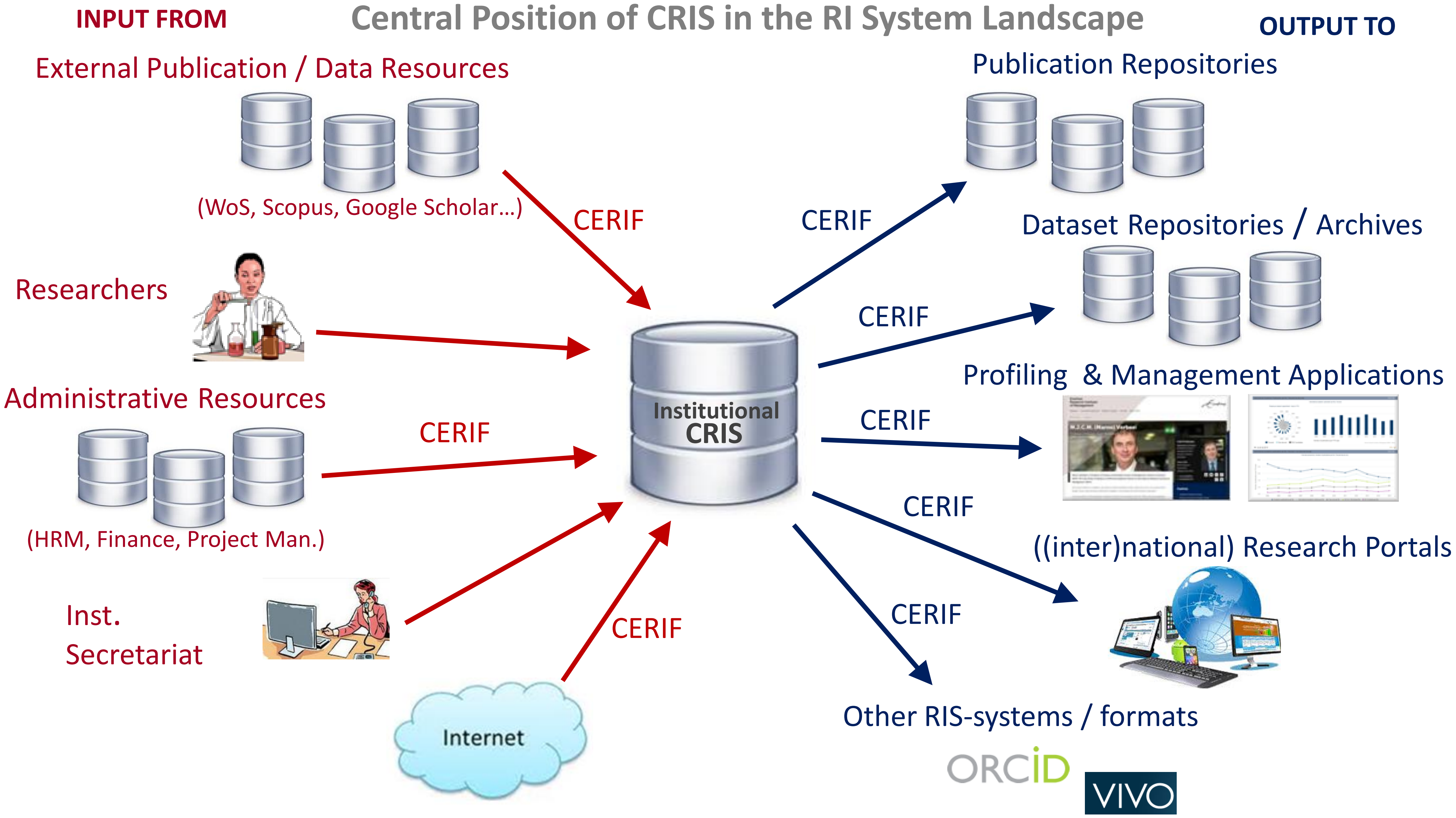
The NIH-Funded Research Workforce

Data by Gender

"My World"

FAIR Model – Technological Part: THE ROLE OF CRIS's

Central Position of CRIS in the RI System Landscape



INPUT FROM

External Publication / Data Resources

(WoS, Scopus, Google Scholar…)

CERIF

Researchers

Administrative Resources

CERIF

(HRM, Finance, Project Man.)

Inst.
Secretariat

Internet

CERIF

Institutional
CRIS

OUTPUT TO

Publication Repositories

CERIF

Dataset Repositories / Archives

CERIF

Profiling & Management Applications

CERIF

CERIF

((inter)national) Research Portals

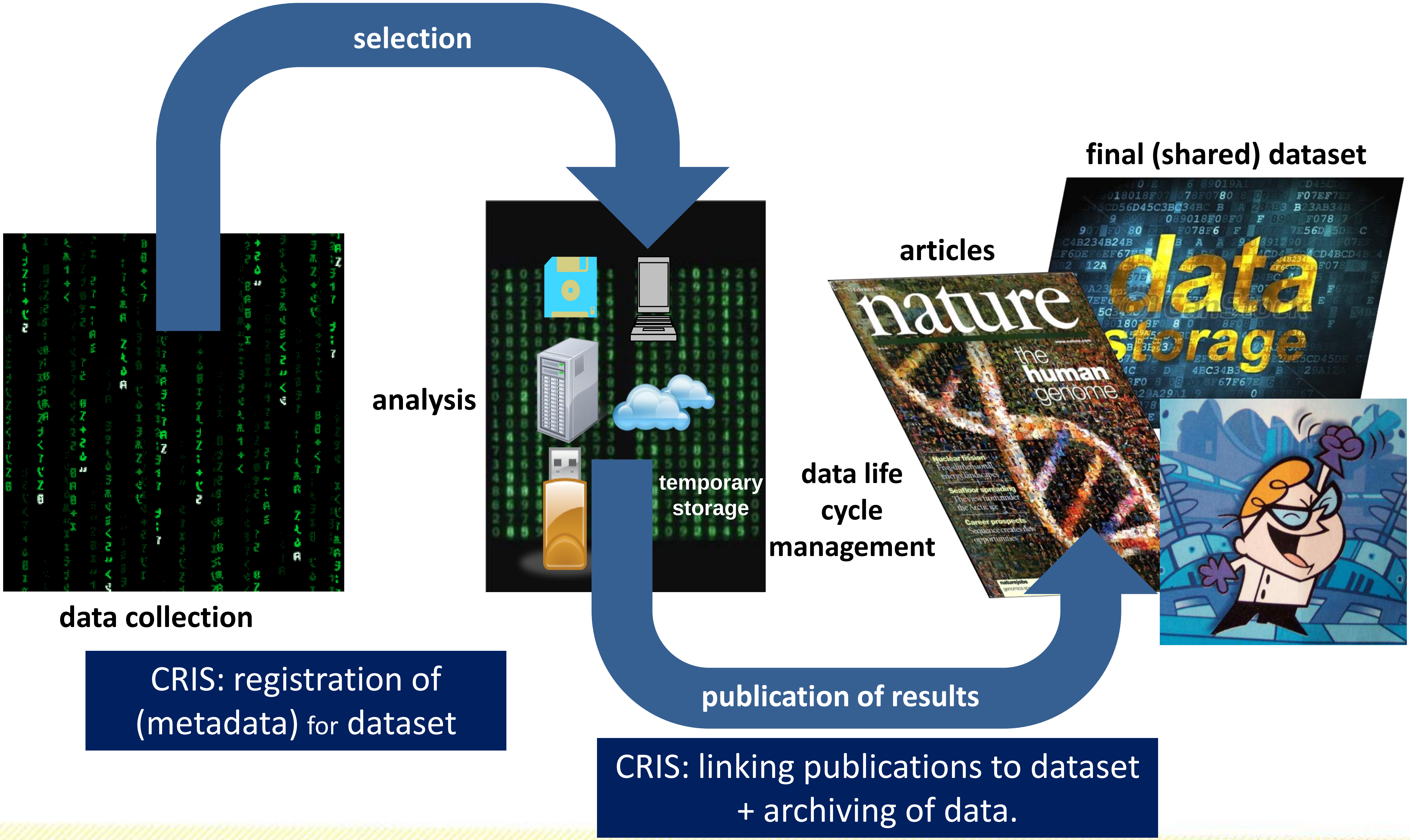
CERIF

Other RIS-systems / formats

ORCID

VIVO

FAIR Model – Technological Part: a look at the DATA LIFE CYCLE

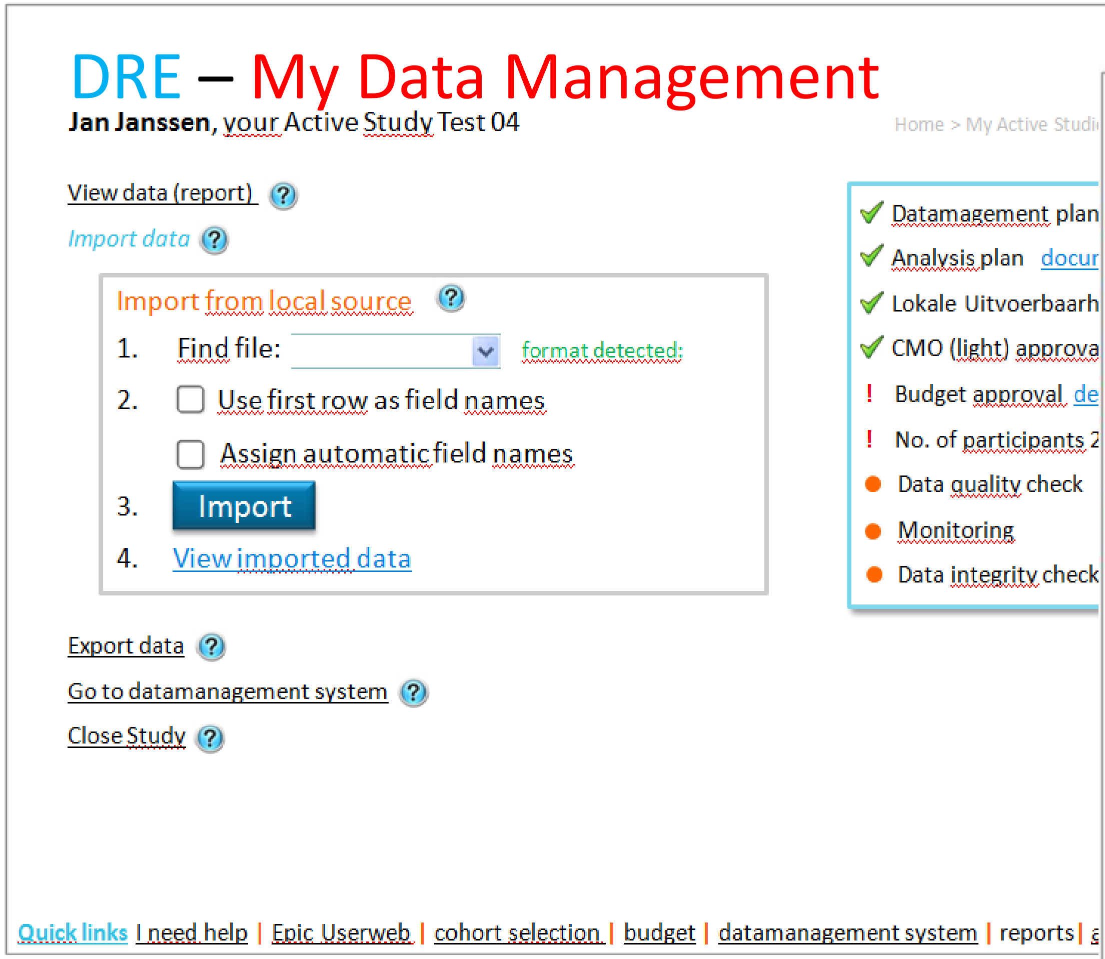


CRIS: registration of (metadata) for dataset

CRIS: linking publications to dataset + archiving of data.

FAIR Model – Technological Part: integrating CRIS-functionality in the Virtual Research Environment of the Researcher (VRE)

A “traditional” VRE may look like this



DRE – My Data Management

Jan Janssen, your Active Study Test 04

Home > My Active Studi

View data (report) ?

Import data ?

Import from local source ?

1. Find file: format detected:
2. □ Use first row as field names
   □ Assign automatic field names
3. Import
4. View imported data

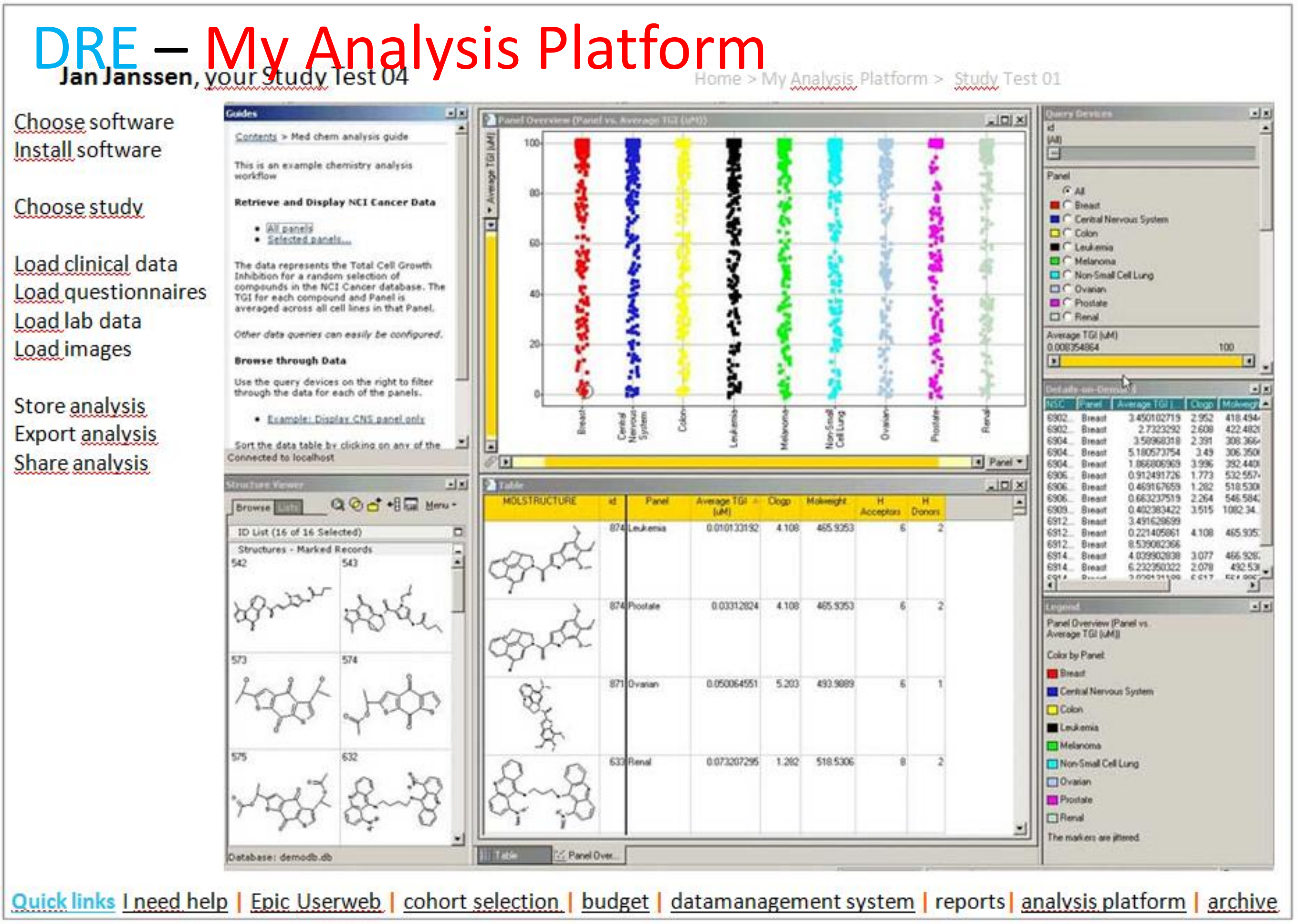
✓ Datamagement plan
✓ Analysis plan docur
✓ Lokale Uitvoerbaarh
✓ CMO (light) approva
! Budget approval de
! No. of participants 2
● Data quality check
● Monitoring
● Data integrity check

Export data ?

Go to datamanagement system ?

Close Study ?

Quick links I need help | Epic Userweb | cohort selection | budget | datamanagement system | reports | a



DRE – My Analysis Platform

Jan Janssen, your Study Test 04

Home > My Analysis Platform > Study Test 01

Choose software
Install software

Choose study

Load clinical data
Load questionnaires
Load lab data
Load images

Store analysis
Export analysis
Share analysis

Guides

Contents > Med chem analysis guide

This is an example chemistry analysis workflow

Retrieve and Display NCI Cancer Data

- All panels
- Selected panels...

The data represents the Total Cell Growth Inhibition for a random selection of compounds in the NCI Cancer database. The TGI for each compound and Panel is averaged across all cell lines in that Panel.

Other data queries can easily be configured.

Browse through Data

Use the query devices on the right to filter through the data for each of the panels.

- Example: Display CNS panel only

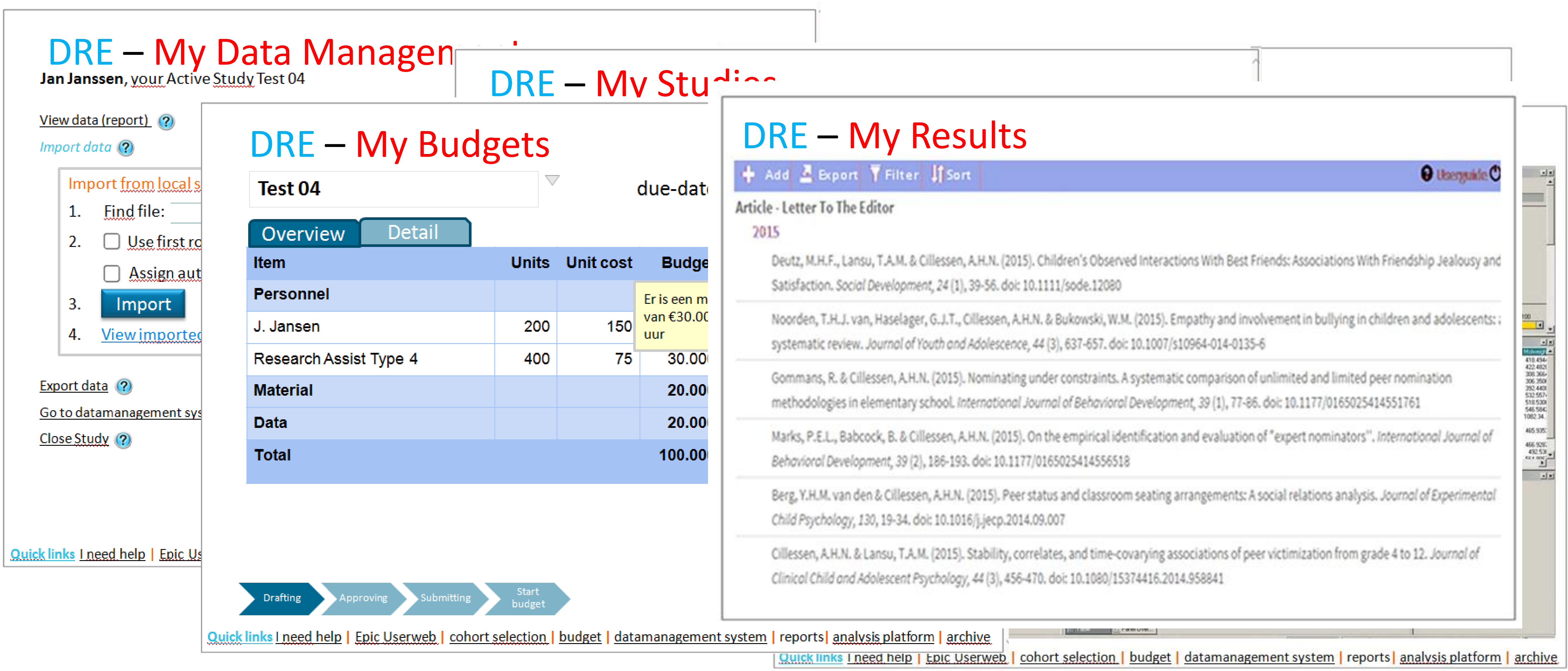
Sort the data table by clicking on any of the

Connected to localhost

Quick links I need help | Epic Userweb | cohort selection | budget | datamanagement system | reports | analysis platform | archive

FAIR Model – Technological Part: integrating CRIS-functionality in the Digital/Virtual Research Environment of the Researcher (VRE)

A “CRIS-integrated” VRE may look like this.



DRE – My Data Managen

Jan Janssen, your Active Study Test 04

View data (report)

Import data

Import from local s

1. Find file:
2. □ Use first ro

   □ Assign aut
3. Import
4. View importec

Export data

Go to datamanagement sys

Close Study

Quick links I need help | Epic Us

DRE – My Studies

DRE – My Budgets

Test 04 due-dat

Overview Detail

Item	Units	Unit cost	Budge
Personnel			
J. Jansen	200	150	
Research Assist Type 4	400	75	30.00
Material			**20.00**
Data			**20.00**
Total			**100.00**

Er is een m
van €30.00
uur

Drafting → Approving → Submitting → Start budget

Quick links I need help | Epic Userweb | cohort selection | budget | datamanagement system | reports| analysis platform | archive

DRE – My Results

Add Export Filter Sort Userguide

Article - Letter To The Editor

2015

Deutz, M.H.F., Lansu, T.A.M. & Cillessen, A.H.N. (2015). Children's Observed Interactions With Best Friends: Associations With Friendship Jealousy and Satisfaction. *Social Development, 24* (1), 39-56. doi: 10.1111/sode.12080

Noorden, T.H.J. van, Haselager, G.J.T., Cillessen, A.H.N. & Bukowski, W.M. (2015). Empathy and involvement in bullying in children and adolescents: a systematic review. *Journal of Youth and Adolescence, 44* (3), 637-657. doi: 10.1007/s10964-014-0135-6

Gommans, R. & Cillessen, A.H.N. (2015). Nominating under constraints. A systematic comparison of unlimited and limited peer nomination methodologies in elementary school. *International Journal of Behavioral Development, 39* (1), 77-86. doi: 10.1177/0165025414551761

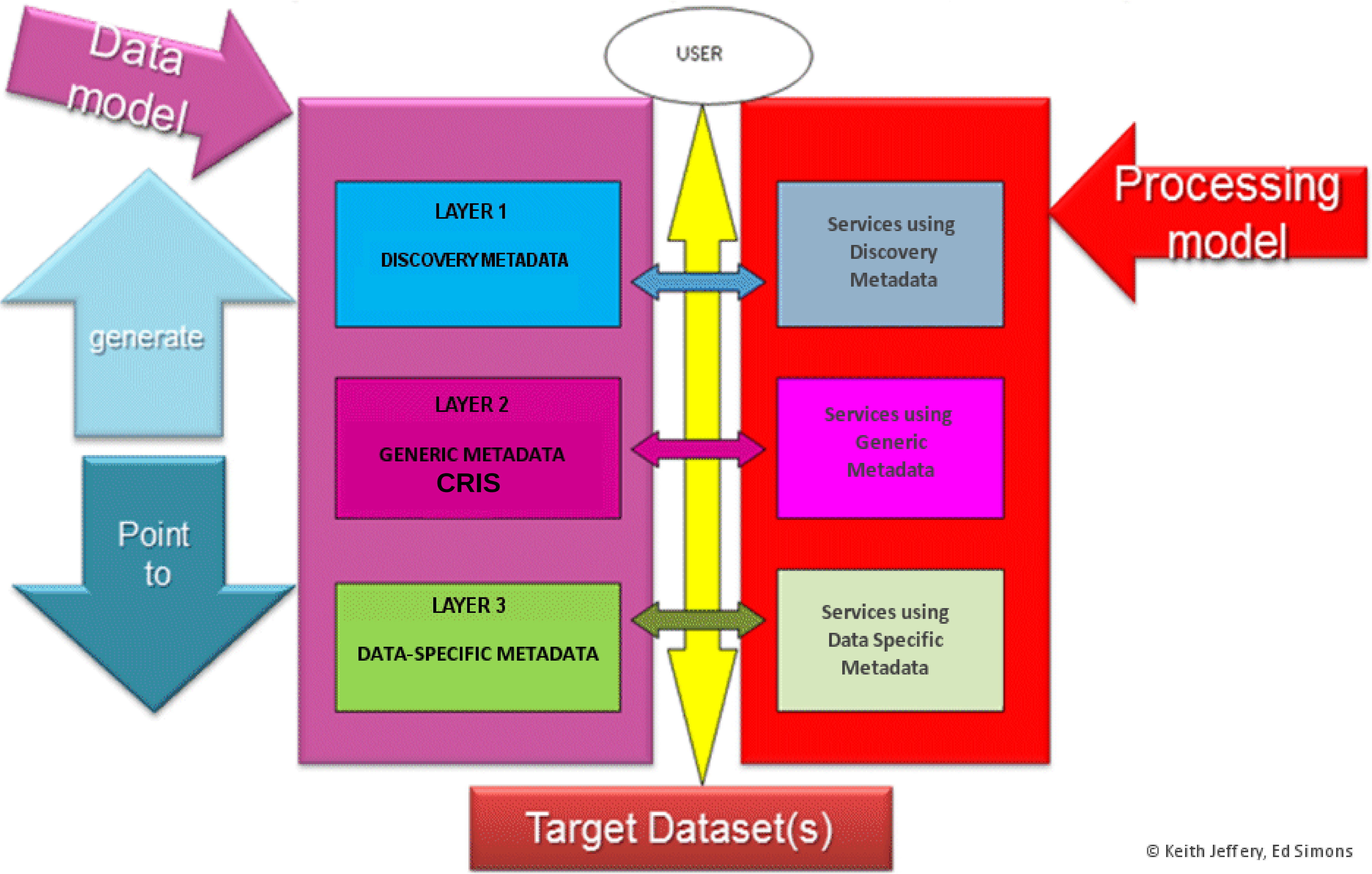
Marks, P.E.L., Babcock, B. & Cillessen, A.H.N. (2015). On the empirical identification and evaluation of "expert nominators". *International Journal of Behavioral Development, 39* (2), 186-193. doi: 10.1177/0165025414556518

Berg, Y.H.M. van den & Cillessen, A.H.N. (2015). Peer status and classroom seating arrangements: A social relations analysis. *Journal of Experimental Child Psychology, 130*, 19-34. doi: 10.1016/j.jecp.2014.09.007

Cillessen, A.H.N. & Lansu, T.A.M. (2015). Stability, correlates, and time-covarying associations of peer victimization from grade 4 to 12. *Journal of Clinical Child and Adolescent Psychology, 44* (3), 456-470. doi: 10.1080/15374416.2014.958841

Quick links I need help | Epic Userweb | cohort selection | budget | datamanagement system | reports| analysis platform | archive

FAIR Model – Technological Part: METADATA LAYER STRUCTURE



© Keith Jeffery, Ed Simons

Metadata Layer Structure: A CONCRETE EXAMPLE

- Suppose: Marine Biology research into the water quality in a certain region of the Indian Ocean and its effects the Coral Reefs in that region.
- Water samples are taken regularly in various parts of this region.
- In layer 2, the central generic layer (CRIS), metadata are stored such as:
  - Unique identifier of the dataset (e.g. DOI)
  - Name/title of the dataset
  - Language
  - Conditions for access and re-use
  - Possible restrictions for public access
  - Institute which conducted the research
  - (names, titles, roles, etc… of) Researchers involved
  - Responsible person and contact person for the dataset
  - Project as part of which the research was carried out
  - Publications based upon/linked to the dataset
  - Geographical coordinates of the of the Indian Ocean region the research applied to
  - Pointer (URI) to the data-specific metadata in layer 3, etc…

Metadata Layer Structure: A CONCRETE EXAMPLE

- From layer 2 the discovery metadata (e.g. DC) for layer 1 are automatically generated and “pushed” into the data repository (or either directly exposed on the Internet).

- In layer 3 the discipline- or data-specific metadata are stored, e.g. the geographical coordinates of the specific parts of the region, specific classification schemes of the coral flora and fauna, chemical analysis procedures applied, equipment used etc…

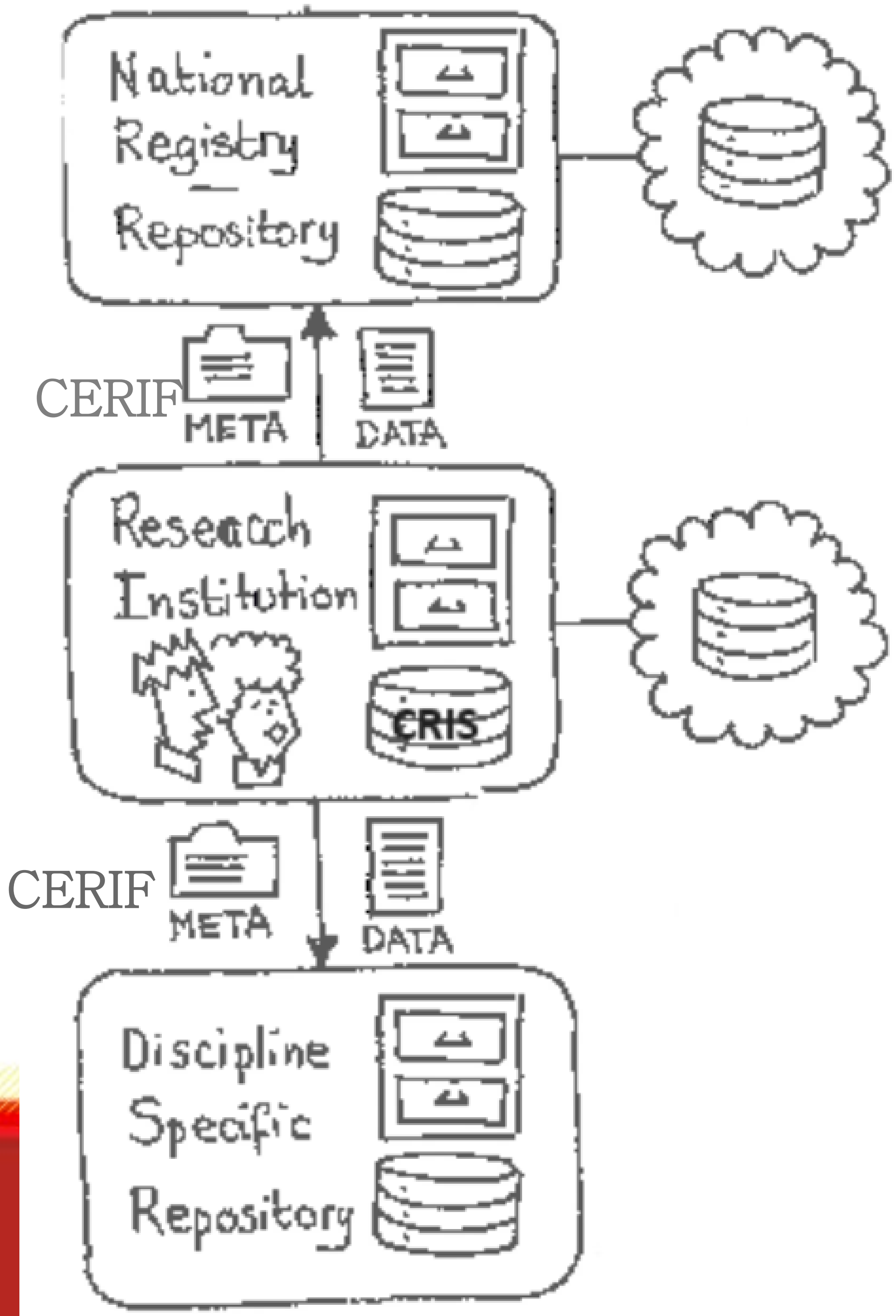
- For each layer specific services may be implemented, e.g.: harvesting services on layer 1, dataset profile creation or linked data services on layer 2, integrity control services on layer 3, etc…

FAIR Model – The Service Part: LEVELS and PLAYERS

- There are four levels or "players" included in the FAIR service model:

  - The *local or institutional* level (institutional repository/CRIS)
  - The *national* level (national registry/repository)
  - The d*iscipline* level (disciplinary registry/repository)
  - The *international* level (international repository).

FAIR Model – Service Part: LEVELS and PLAYERS



THE SERVICE MODEL: LAYERS and PLAYERS

- The *national / international registry* is the obvious party to:
  - register metadata to the global search engines;
  - offer a national repository service (ideally via cloud storage);
  - implement and provide the request-approve-receive workflow (see further);
  - assist research institutions that choose to host (some of) their datasets on premise or in the cloud.

- For whatever reason some research institutions may choose to host (some of) their datasets on premise or in the cloud themselves, others will store their datasets in the national repository.

Discipline Specific Repositories

- Researchers working in a discipline with a discipline specific repository (such as astrophysics and genomics) will in general access this repository directly, both for registering their own datasets, and searching for other researcher’s datasets.

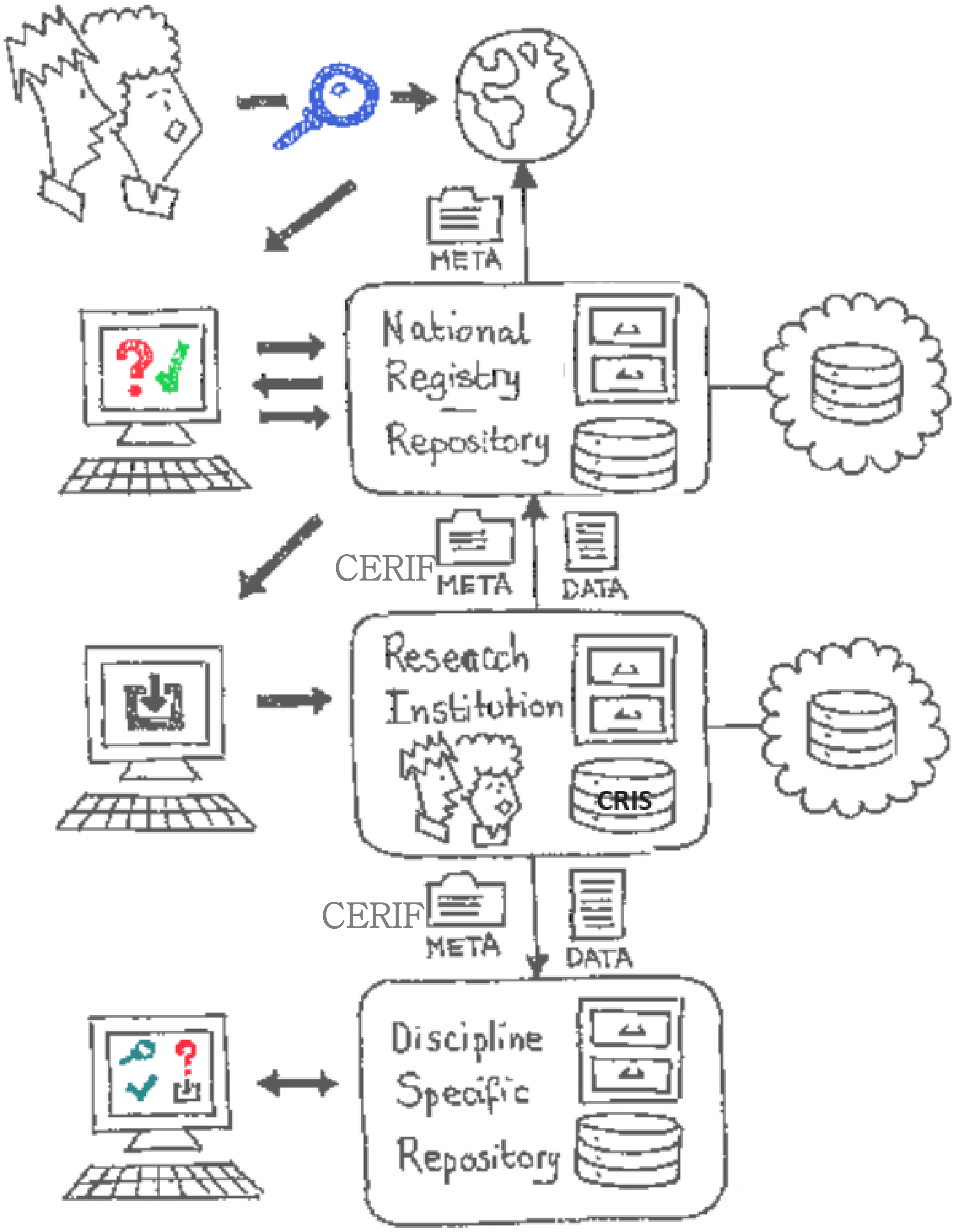
- Such repositories offer elaborate, discipline specific search capabilities, access request and granting workflows, and download functionality.

- For the purpose of completeness it is desirable that national registries hold the (generic) metadata of datasets in discipline specific repositories. To this end the research institution should assist their researchers (through some automated process) to upload the generic metadata to the national registry without requiring double work from the researcher.

IMPLEMENTING THE “FAIR DATA INFORMATION MODEL”

- To implement FAIR infrastructures in a scalable way, we need to:
  - Implement the following functions: *Find, Request, Approve, Receive*.
  - Ensure that *datasets are stored safely, and remain accessible.*
  - Ensure that the *metadata are correct, standardized and detailed and complete enough.*

IMPLEMENTING THE FAIR DATA MODEL: integrating the technological and service part



CERIF

CRIS

CERIF

Example

- A researcher searches for a dataset based on certain meta data.
- The search engine produces a reference to a dataset registered in a national registry. The dataset itself is hosted in some research institution.
- The researcher issues a request for the dataset via the national registry services interface.
- The national registry forwards the request to the research institution hosting the dataset.
- If the request is granted, the national registry communicates the request specific download link to the researcher.
- The researcher accesses the research insitution repository to download the dataset. (In case the dataset resides on the national repository, the registry provides the appropriate link as well.)

- Discipline specific repositories generally provide all four steps, and the searching capabilities are much more discipline specific.

ADDING USER SUPPORT: The Front Office-Back Office Model

- To make the model work, an optimal user support structure is necessary. An example of such a structure is the so-called

“Front Office-Back Office” model

ADDING USER SUPPORT: The Front Office-Back Office Model

- The Front Office is located on the local (institutional) level, *with the library being an obvious location for it*, close to the researcher and takes care of the 1st line support.

  This includes a.o.t.:

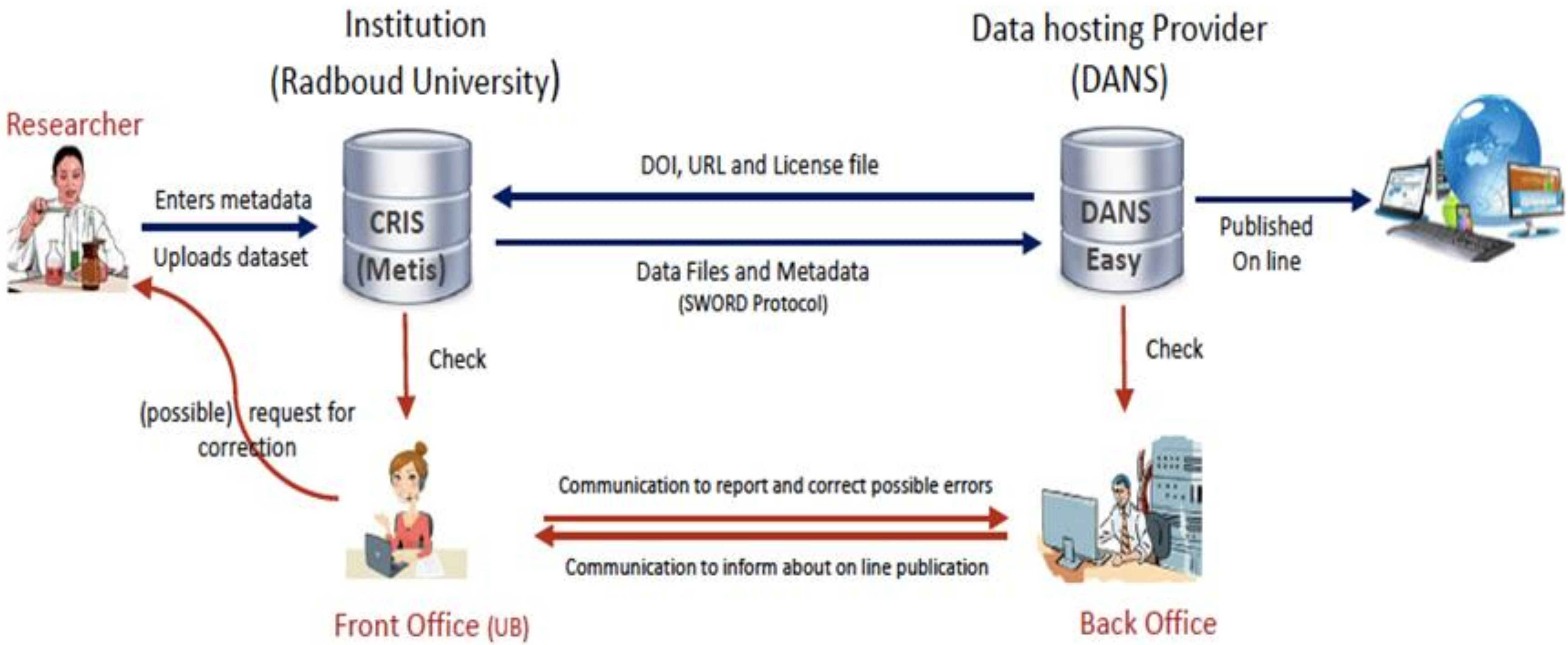
  - Help desk for user support of the dataset registration , e.g.
    - how to use the registration interface,
    - how to structure the data file set according to the requirements regulations of the data hosting organisation,
  - Working out multimedia demo''s as well giving hands-on triaingen to an insitutes researhchers.
  - Check on completeness and correct structure and format of the dataset files
  - Help in writing data management plans.
  - To insitutes: support in developing / formulating a data management policy.
  - Development and management of an information web site.
  - etc…

ADDING USER SUPPORT TO THE MODEL: The Front Office-Back Office Model

- The “Back-Office” functionality is located at the (inter)national or disciplinary level (data hosting services level) and concerns 2nd level support to and through the Front Office.

ADDING USER SUPPORT: The Front Office-Back Office Model

- (with an example from The Netherlands currently being implemented)



SUMMARY: “FAIR DATA INFORMATION INFRASTRUCTURES”

- The model concerns the “FAIR” aspect of data, meaning it deals with information (metadata) promoting and supporting *the findability, interoperability, interpretability and reusability* of datasets.

- The model consists of two parts:
  - *A technological solution.*
  - *A service and support solution*

- *CRIS systems are core elements of the technological solution* since they provide rich additional metadata on datasets and put the datasets and their metadata into their proper context, and so significantly enhance the *FAIR-ness* of datasets.

- *Another crucial element* underlying the technological solution *is the 3-layer metadata model* developed by euroCRIS.

SUMMARY: “FAIR DATA INFORMATION INFRASTRUCTURES”

- *Exchange of information* between parts and levels of the model *should be based on a standard. CERIF is an obvious choice for this*.

- To make the infrastructure and the (FAIR) services built on it work, *organisations on the local, national, international and disciplinary level are involved.*

- *An optimal support organisation, for which the “Front Office – Back Office” is an appropriate model is of the utmost importance*. Especially the Front Office function is vital to make the model work for researchers and institutes.

- Last but not least: the *“FAIR Data Information Infrastructure” should be an integrated part of national and international Research Data Infrastructures* (such as the EU Open Science Cloud), in order for these infrastructures to work optimally