

# Creating a Joint Metadata Domain for the Social Sciences and Humanities in Europe

Catharina Wasner  
GESIS – Leibniz Institute for the Social Sciences

Stephanie Roth, Olof Olsson  
SND - Swedish National Data Service

EDDI 2014, December 2, London

## DASISH Task 5.4

- Goal: “an integrated metadata domain for the SSH disciplines” (DASISH DoW, p. 22)
- Project partners:
  - **MPI-PL** – Max Planck Institute for Psycholinguistics: task coordination/technical infrastructure
  - **GESIS** – Leibniz Institute for the Social Sciences
  - **SND** – Swedish National Data Service
  - **ICLTT** – Institute for Corpus Linguistics and Text Technology
- Main contributors:  
**Daan Broeder**, **Matej Durco**, **Binyam Gebrekidan Gebre**,  
**Przemyslaw Lenkiewicz**, **Kees Jan van de Looij**, **Olof Olsson**,  
**Stephanie Roth**, **Catharina Wasner**, **Bartholemeus Worcslav**

# Requirements

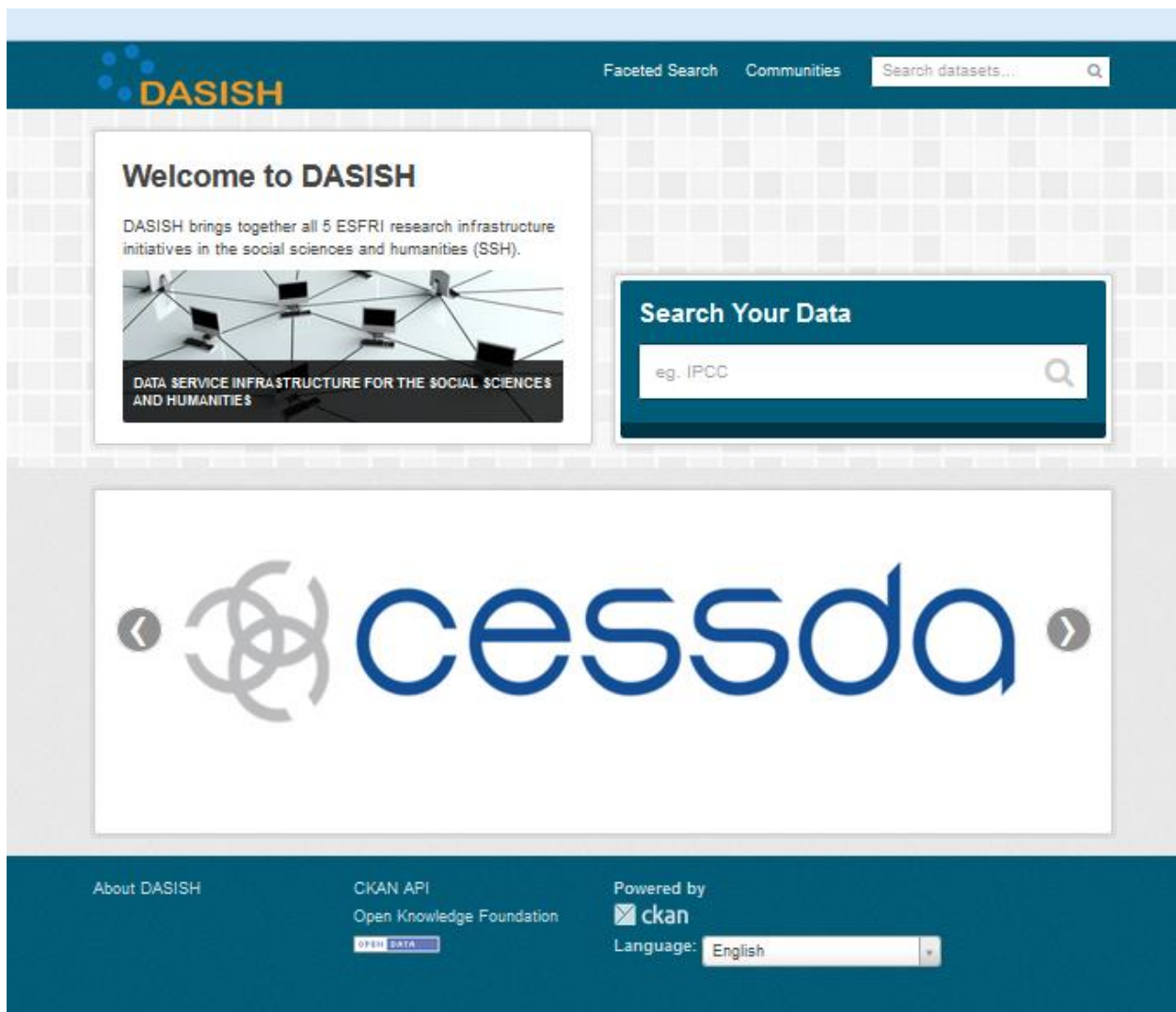
The DASISH catalogue should

- include interdisciplinary metadata from the Social Sciences and Humanities (SSH) - DARIAH, CLARIN, CESSDA - and
- make use of smart search techniques such as faceted browsing.

# The Use of Interdisciplinary Metadata Catalogues

- Enhance the visibility of SSH research data
- Show metadata from different sources via a single application
- Provide search for several disciplines and support cross-disciplinary research interests
- Promote a culture of data sharing, data reuse, verification and citation

Demo <http://ckan.dasish.eu/>



The screenshot shows the DASISH CKAN portal interface. At the top, there is a dark blue header with the DASISH logo on the left, and navigation links for 'Faceted Search' and 'Communities' in the center. On the right of the header is a search bar labeled 'Search datasets...' with a magnifying glass icon. Below the header, the main content area has a light gray background with a subtle grid pattern. On the left, a white box titled 'Welcome to DASISH' contains a paragraph about the infrastructure and a small image of a network of laptops. To the right of this box is a dark blue search box titled 'Search Your Data' with a search bar containing the text 'eg. IPCC' and a magnifying glass icon. Below these elements is a large white box featuring the CESSDA logo, which consists of a stylized circular icon and the text 'cessda' in blue. At the bottom, a dark blue footer contains links for 'About DASISH', 'CKAN API', and 'Open Knowledge Foundation'. On the right side of the footer, it says 'Powered by' followed by the CKAN logo and a 'Language' dropdown menu set to 'English'. In the bottom right corner, there is a logo for the 'SEVENTH FRAMEWORK PROGRAMME'.

**Welcome to DASISH**

DASISH brings together all 5 ESFRI research infrastructure initiatives in the social sciences and humanities (SSH).

**Search Your Data**

eg. IPCC

**cessda**

About DASISH

CKAN API  
Open Knowledge Foundation  
OPEN DATA

Powered by  
ckan  
Language: English

SEVENTH FRAMEWORK PROGRAMME

# Workflow

1. Collecting OAI end-points
2. Choosing a catalogue software
3. Deciding on the set of facets and fields
4. Creating suitable mappings
5. Metadata harvesting
6. Metadata quality improvement
7. Metadata normalization

# Workflow

- 1. Collecting OAI end-points**
2. Choosing a catalogue software
3. Deciding on the set of facets and fields
4. Creating suitable mappings
5. Metadata harvesting
6. Metadata quality improvement
7. Metadata normalization

## 1. Collecting OAI end-points

- Questions:
  - Harvesting metadata directly from the metadata provider or harvesting other harvesters?
  - Where do we get lists of metadata providers (preferably OAI-PMH) from?
- Results:
  - 8 metadata providers (CESSDA)
  - 20 metadata providers (CLARIN)
  - 25 metadata providers (DARIAH)
  - Total: 53 metadata providers



# Workflow

- 1. Collecting OAI end-points**
2. Choosing a catalogue software
3. Deciding on the set of facets and fields
4. Creating suitable mappings
5. Metadata harvesting
6. Metadata quality improvement
7. Metadata normalization

# Workflow

1. Collecting OAI end-points
- 2. Choosing a catalogue software**
3. Deciding on the set of facets and fields
4. Creating suitable mappings
5. Metadata harvesting
6. Metadata quality improvement
7. Metadata normalization

## 2. Choosing a catalogue software

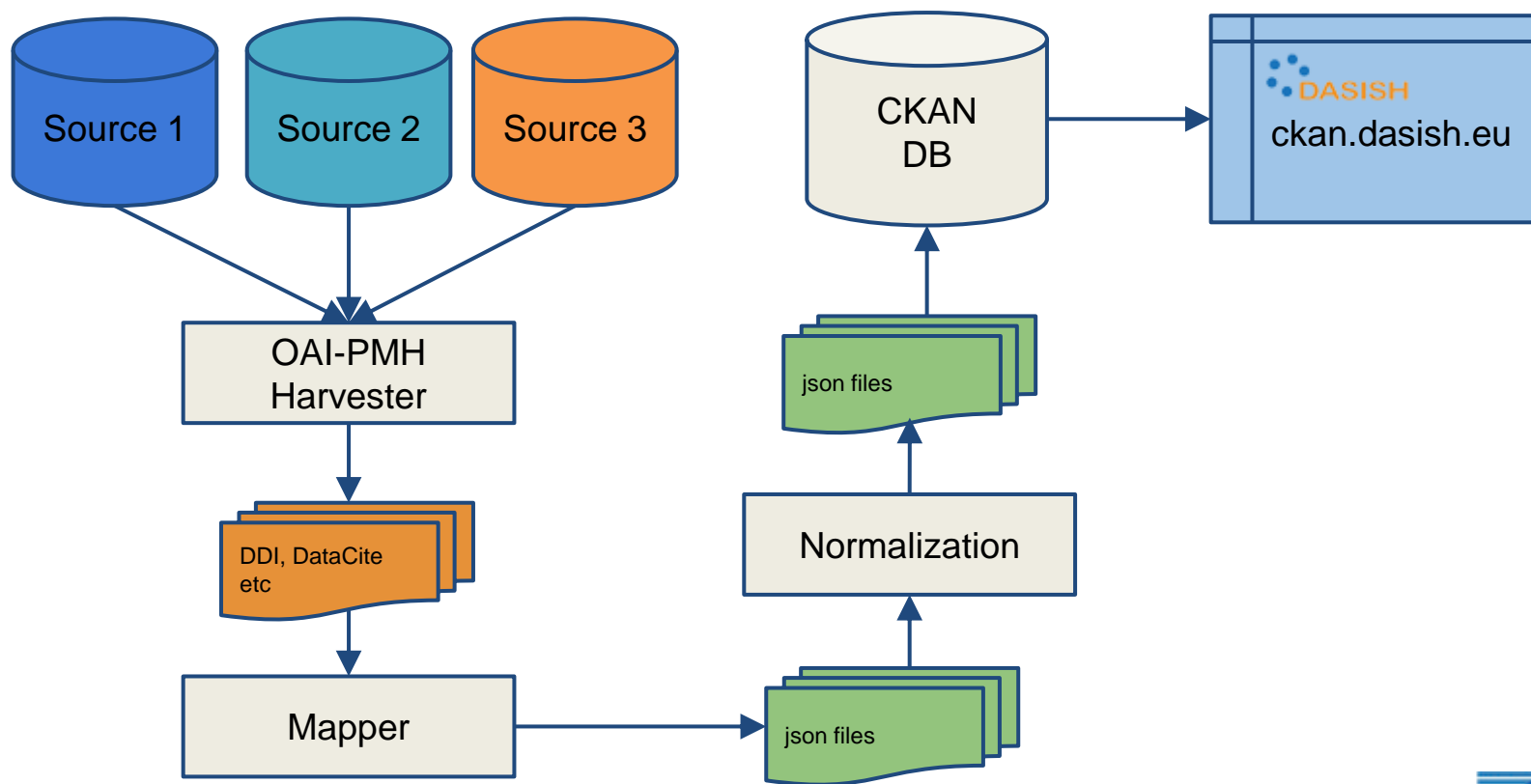
# CKAN

- Developed by the open knowledge foundation
- publish and manage datasets (import functions)
- search and discovery
- harvesting (pull in data from existing repos)
- store raw data and metadata (for each dataset)
- visualise data, advanced geospatial features
- community: engage with users and others
- API: customise and extend
- see <http://ckan.org/features/> for more information

# Choice of ckan for the metadata portal

- CKAN have been used earlier by MPI and is built on proven technology
- API to import metadata via json/csv
- Customizable interface

# Technical flow



# Workflow

1. Collecting OAI end-points
- 2. Choosing a catalogue software**
3. Deciding on the set of facets and fields
4. Creating suitable mappings
5. Metadata harvesting
6. Metadata quality improvement
7. Metadata normalization

# Workflow

1. Collecting OAI end-points
2. Choosing a catalogue software
- 3. Deciding on the set of facets and fields**
4. Creating suitable mappings
5. Metadata harvesting
6. Metadata quality improvement
7. Metadata normalization

### 3. Deciding on the set of facets and fields

## Search Facets

- |                     |                |
|---------------------|----------------|
| 1. Communities      | 7. Country     |
| 2. Creator          | 8. Collection  |
| 3. Language         | 9. Discipline  |
| 4. Creation Year    | 10. Subject    |
| 5. Publication Year | 11. OAI Origin |
| 6. Data Provider    |                |



### Communities

DARIAH (302164)

CLARIN (160613)

CESSDA (49894)

### Creator

not applicable (11014)

Tomoko Ishizuka (3154)

Kurt Kreppner (2914)

Sotaro Kita & Mihok... (1899)

Amanda Brown (1627)

[Show More Creator](#)

### Language

French (58139)

German (53547)

English (51312)

Dutch (49041)



**512,671 datasets found**

Order by: Relevance



**Saudargas Paulius**

**Narkevič Jaroslav**

**Prezidento valstybinė našlių renta ir valstybės turto valdymas, 2010 m. spalio**

Tyrimo tikslas: nustatyti Lietuvos gyventojų požiūrį į valstybinės Prezidento našlės rentos skyrimą K. Brazauskienei bei iširti nuomonę apie galimus skirtingus valstybės...

**Rutkelytė Rūta**

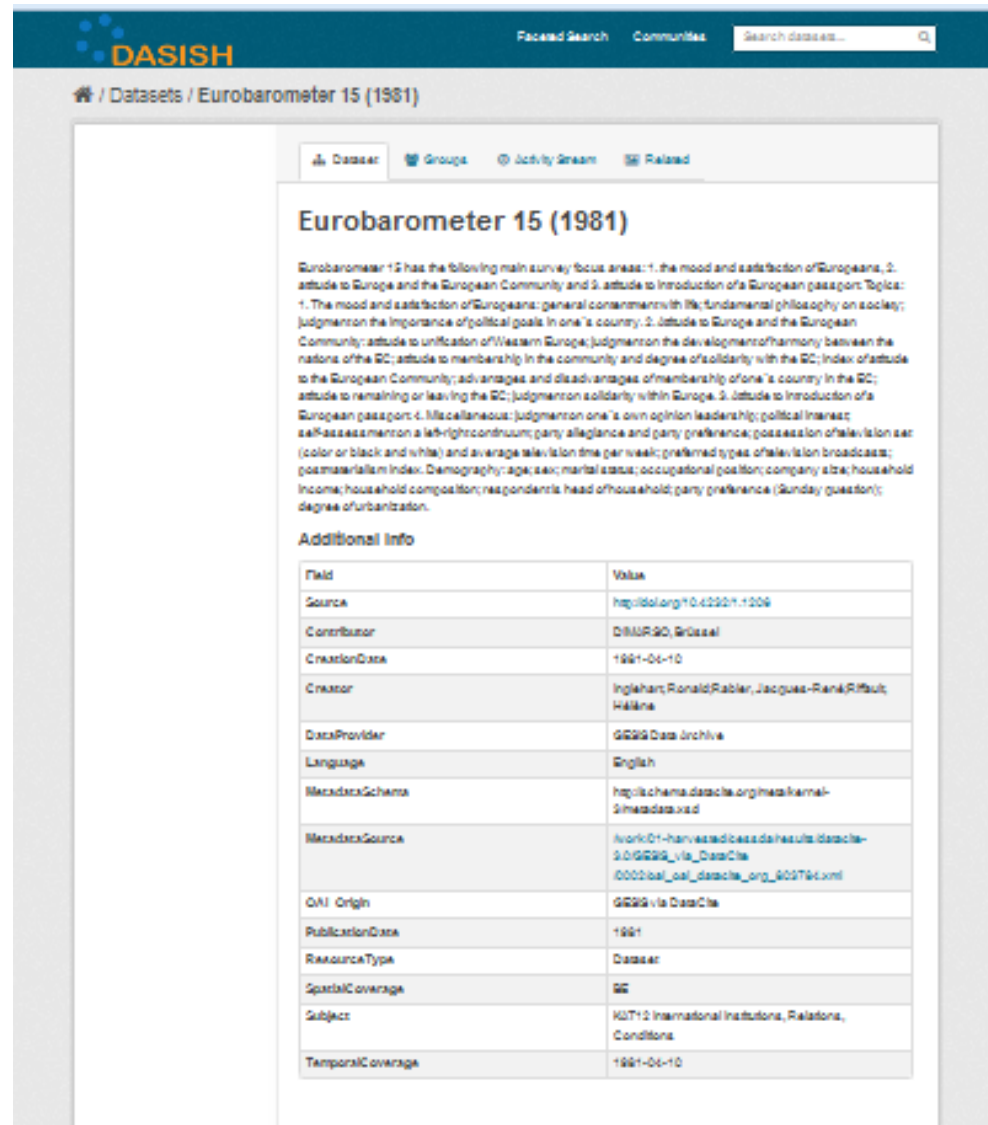
**Atominės elektrinės statyba, 2008 m. gegužė**

Tyrimo tikslas: išsiaiškinti Lietuvos gyventojų ketinimus pirkti naujosios atominės elektrinės akcijų. Pagrindiniai nagrinėti klausimai: kai kuriems ekspertams ir politikams...

# Other Metadata Fields

- Spatial coverage
- Temporal coverage
- Contributor
- Metadata schema
- Metadata source
- Resource type
- Rights
- Data format

<http://ckan.dasish.eu/dataset/ef0c1e882b57c80f542f6021110036009bb5de864ecec9e5fb05751ebd3e5de3>



The screenshot shows the DASISH dataset page for Eurobarometer 15 (1981). The page includes a search bar, navigation tabs (Dataset, Groups, Activity Stream, Related), and a detailed description of the dataset. Below the description is an 'Additional Info' table with the following fields and values:

Field	Value
Source	<a href="http://doi.org/10.4232/1.1206">http://doi.org/10.4232/1.1206</a>
Contributor	DIJURGO, Brussels
CreationDate	1981-04-10
Creator	Inglehart, Ronald; Rabier, Jacques; Rank, Riffault, Hélène
DataProvider	GESIS Data Archive
Language	English
MetadataSchema	<a href="http://ckan.dasish.eu/metadata/schema-1.0/ckan_dasish_v1.0.xsd">http://ckan.dasish.eu/metadata/schema-1.0/ckan_dasish_v1.0.xsd</a>
MetadataSource	<a href="http://ckan.dasish.eu/metadata/schema-1.0/ckan_dasish_v1.0.xsd">http://ckan.dasish.eu/metadata/schema-1.0/ckan_dasish_v1.0.xsd</a>
DOI: Origin	GESIS via DataCite
PublicationDate	1981
ResourceType	Dataset
SpatialCoverage	EEC
Subject	K0112 International Institutions, Relations, Conditions
TemporalCoverage	1981-04-10

# Workflow

1. Collecting OAI end-points
2. Choosing a catalogue software
- 3. Deciding on the set of facets and fields**
4. Creating suitable mappings
5. Metadata harvesting
6. Metadata quality improvement
7. Metadata normalization

# Workflow

1. Collecting OAI end-points
2. Choosing a catalogue software
3. Deciding on the set of facets and fields
- 4. Creating suitable mappings**
5. Metadata harvesting
6. Metadata quality improvement
7. Metadata normalization

## 4. Creating suitable mappings

Format harvested

- DDI 1.2.2 (nesstar)
- DDI 3.1
- DataCite
- CMDI (component metadata infrastructure) -Clarín
- Dublin Core

Mappings available on:

<https://github.com/DASISH/md-mapping>

# Mapping via xpath

```
<field name="title">
  <xpath>
    /ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title[@xml:lang='en']/text()
  </xpath>
  <xpath>/ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/text()</xpath>
  <xpath>
    /ddi:DDIInstance/s:StudyUnit/r:Citation/dc:DCElements/
    dc2:title[@xml:lang='en']/text()
  </xpath>
</field>
```

# Mapping problems

- Values in certain fields are not always what you think. Mixed content in identifier fields (url:s, doi, internal identifiers, shortnames etc.)
- xml:lang is not always used in some communities
- Hard to get Language code from DDI (DDI-C and DDI-L) if its event there at all.
- More complex logic could have been expressed using XSLT

# Workflow

1. Collecting OAI end-points
2. Choosing a catalogue software
3. Deciding on the set of facets and fields
- 4. Creating suitable mappings**
5. Metadata harvesting
6. Metadata quality improvement
7. Metadata normalization



# Workflow

1. Collecting OAI end-points
2. Choosing a catalogue software
3. Deciding on the set of facets and fields
4. Creating suitable mappings
- 5. Metadata harvesting**
6. Metadata quality improvement
7. Metadata normalization

## 5. Metadata harvesting

- Harvesting done via OAI-PMH
  - CESSDA
    - harvested from **7** out of **9** providers
    - 49,894 records
  - CLARIN
    - harvested from **4** out of **20** providers
    - 160,613 records
  - DARIAH
    - harvested from **14** out of **25** providers
    - 302,164 records

Tool developed by MPI:

<https://github.com/TheLanguageArchive/md-mapper>

# Workflow

1. Collecting OAI end-points
2. Choosing a catalogue software
3. Deciding on the set of facets and fields
4. Creating suitable mappings
- 5. Metadata harvesting**
6. Metadata quality improvement
7. Metadata normalization

# Workflow

1. Collecting OAI end-points
2. Choosing a catalogue software
3. Deciding on the set of facets and fields
4. Creating suitable mappings
5. Metadata harvesting
- 6. Metadata quality improvement**
7. Metadata normalization

## 6. Metadata quality improvement

- Different date fields (Temporal Coverage, Creation Date, Publication Date) to increase the conformance of user expectations
- „If-then-else“ logical constructions to avoid data sparseness (e.g. if no creator then use project/organization)
- Normalization to overcome differences in controlled vocabularies

# Workflow

1. Collecting OAI end-points
2. Choosing a catalogue software
3. Deciding on the set of facets and fields
4. Creating suitable mappings
5. Metadata harvesting
- 6. Metadata quality improvement**
7. Metadata normalization

# Workflow

1. Collecting OAI end-points
2. Choosing a catalogue software
3. Deciding on the set of facets and fields
4. Creating suitable mappings
5. Metadata harvesting
6. Metadata quality improvement
- 7. Metadata normalization**

**Language**

eng (148662)

nl (26576)

English (23803)

Dutch (19990)

en (12804)

German (11052)

French (6012)

es (5450)

nld (4894)

*(Language facet before the normalization)*

## 7. Metadata normalization

Dates

convert to yyyy-mm-dd

Country

convert to ISO 3166

Language

convert to ISO 639-3

Scripts available on:

<https://github.com/DASISH/jmd-scripts>



# Workflow

1. Collecting OAI end-points
2. Choosing a catalogue software
3. Deciding on the set of facets and fields
4. Creating suitable mappings
5. Metadata harvesting
6. Metadata quality improvement
- 7. Metadata normalization**

# Future of the DASISH catalogue

- Question:  
How can we keep the DASISH catalogue available in the future since the DASISH project ends 2014?
- Answer:  
Turn the catalogue over to the EUDAT project which is running B2FIND.
- Deliverable available on  
<http://dasish.eu/deliverables/>
- Documentation available on  
<https://github.com/DASISH>

# Thank you!