# Ways for a Machine-actionable Processing Chain for Identifier, Metadata, and Data

Workshop on Metadata and Persistent Identifiers
for Social and Economic Data
May 7-8 2012, Berlin

Joachim Wackerow
GESIS – Leibniz Institute for the Social Sciences

# Outline

- Motivation
- What is a machine-actionable processing chain?
- Use cases
- Current possibilities
- Conclusion

# Motivation

- “Metadata and Persistent Identifiers are complementary ingredients in the world of digital object management, making it possible to find, reuse, reference, cite and link digital content”
- Good metadata is necessary to identify the relevant data when searching, reusing, or linking data.
- A program is as likely to follow a URL as a person is.
- Possible machine-actionable scenarios :
  - Harvesting metadata
  - Processing metadata/data
- Better interoperability between the levels of PID, different types of metadata and the data itself.
- **Compound object of PID, metadata, and data to make something useful.**

# Difference Publication/Data

- Data has a different nature than publication.
- A publication such as an article is for human consumption.
  - A retrieved article can be immediately read.
- Data is usually for processing by programs.
  - Metadata is necessary to understand the data. Programs are required for processing.
  - Example for important metadata
    - Type of file format, i.e. CSV, binary, program-specific
    - Logical file structure, i.e. rectangular,
    - Unit of record, i.e. person
    - Meaning of columns, i.e. age

# What Is a Machine-actionable Processing Chain?

- Processing of any components related to a PID
  - Catalog metadata (like DataCite)
  - Rich metadata (like DDI, SDMX)
  - Data (in different formats)

# Harvesting Metadata

- Building value-added services with registries/portals. Examples:
  - Providing searching possibilities for
    - allowing data to be found by relevant criteria
    - bringing related/similar data together
    - distinguishing dissimilar data
  - Linking data and publications in an integrated way

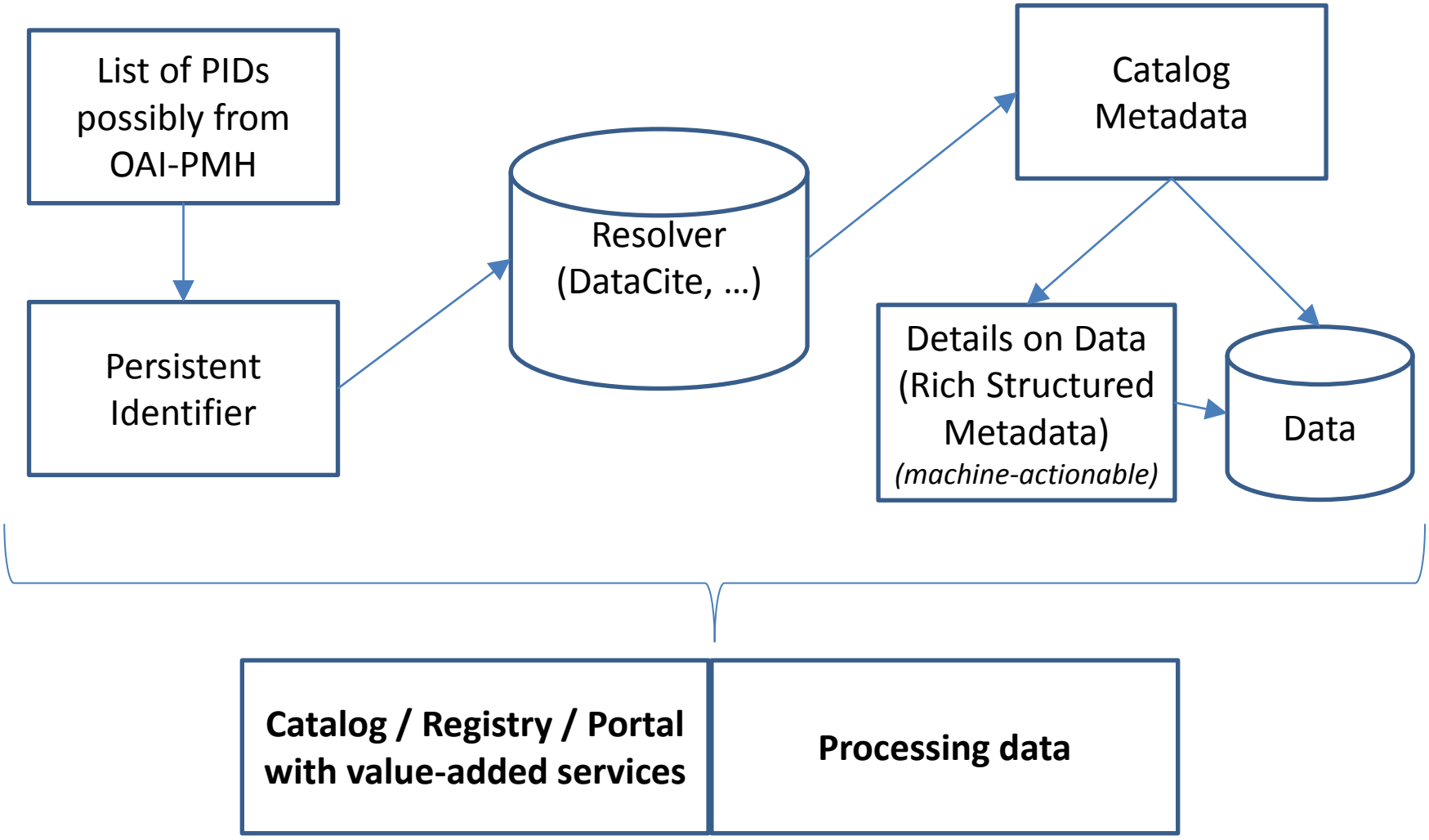
# Use Case: Specific Search

- Search for studies using the educational classification ISCED
  - With catalog metadata: search for subject „education“
    - Result would be broad
  - With rich metadata: search for variable ISCED
    - Result is very specific
- Prerequisite would be a registry/portal which uses catalog metadata and rich metadata

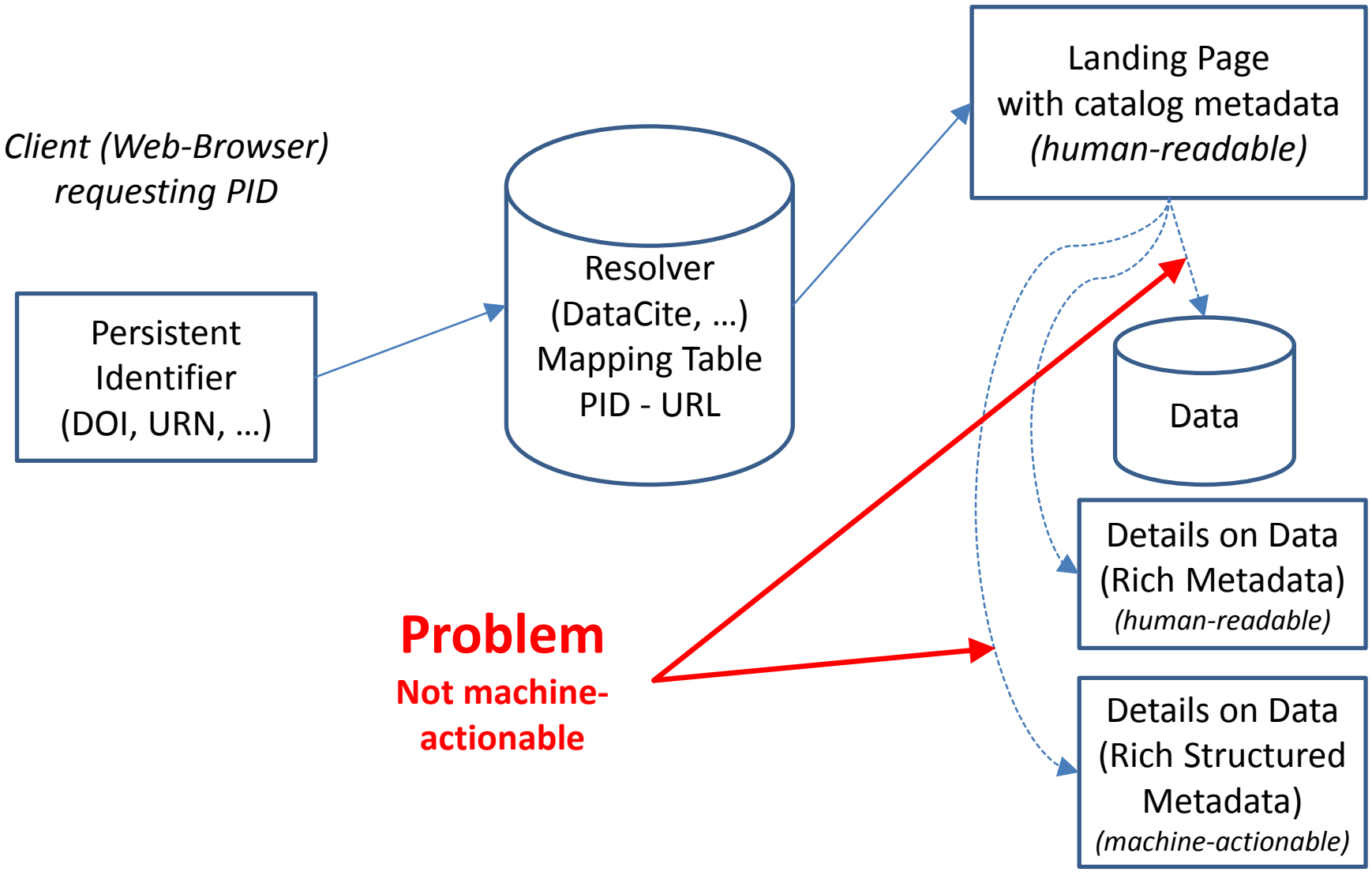
# Processing Data

- Processing data
  - Building subsets
  - Merging data of different sources
  - Analyzing data
- Access to data is often restricted, especially with data on people

# Machine-actionable Path for Value-added Services



# Resolution - Current Status



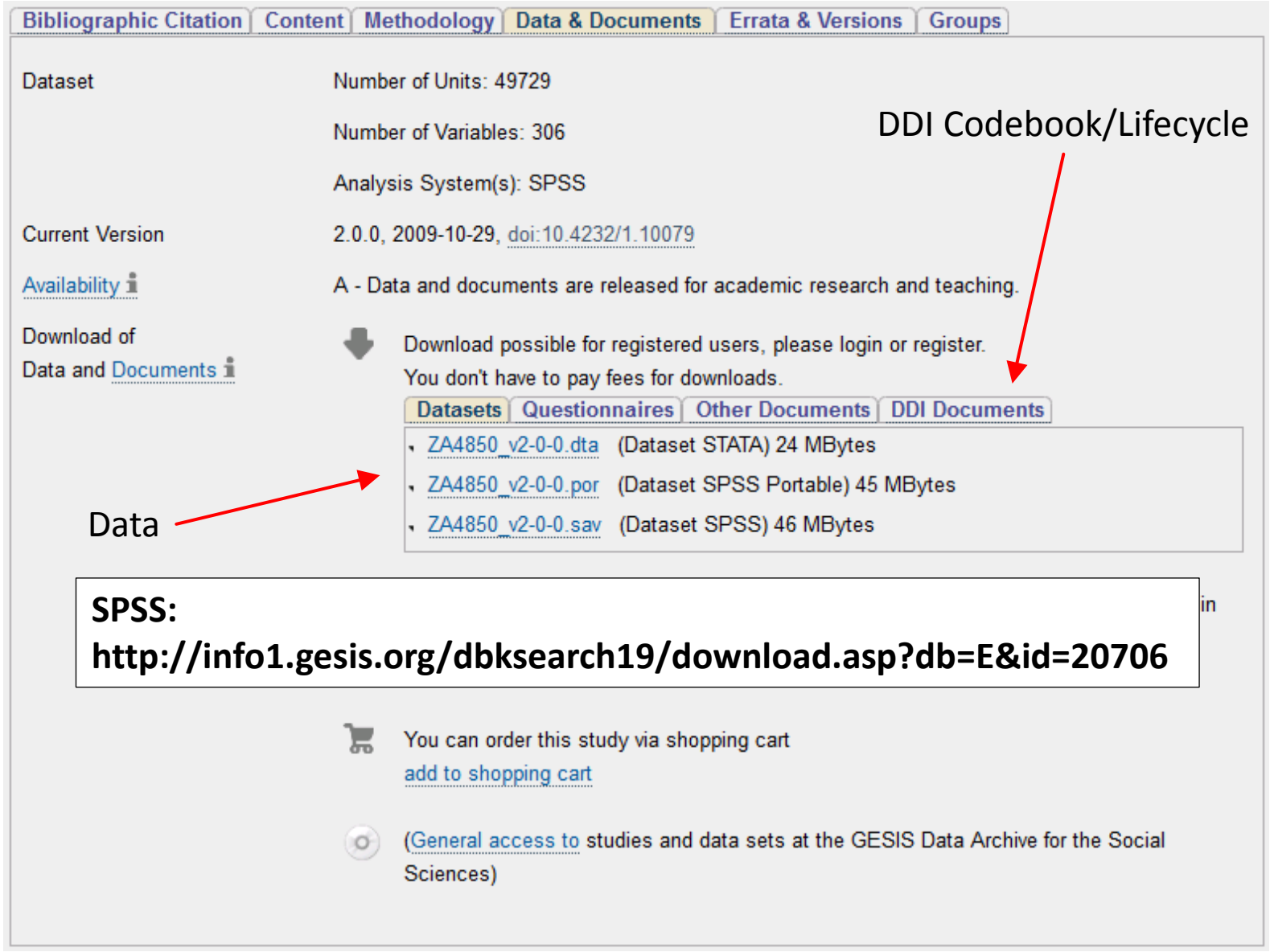
# Current Machine-actionable Possibilities

- APIs of registry agencies to receive machine-actionable content
  - Example: OpenURL
  - Issues: Not scalable, APIs vary from one data provider to next
- Content negotiation
  - Program can resolve a DOI through the standard proxy
    - Desired machine-actionable content can be specified in HTTP Accept header
  - DataCite Content Service (similar to CrossRef)
  - Restrictions: approach cannot be used by web browsers
- HTML Links with direkt link for each representation
  - Issues: Not compliant to web standards like REST

# Current Machine-actionable Possibilities (continued)

- Semantic Web / Linked data
  - Den Haag Manifesto (June 2011): Five steps to connecting Persistent Identifiers and Linked Open Data
    - Most of these steps involve adopting linked data principles including support for content negotiation.
  - DataCite2RDF – Mapping DataCite Metadata Scheme Terms to ontologies
    - Intention: enabling these metadata to be understood programmatically and integrated automatically with similar data from elsewhere.
  - RDF representation of DataCite metadata
- OAI-PMH, Open Archives Initiative Protocol for Metadata Harvesting
  - DataCite has service in beta status

# GESIS DBK Example – Web Page



# Pangaea Example – Web Page

**PANGAEA²**
**Data Publisher for Earth & Environmental Science**

Not logged in (log in or sign up)

Always quote citation when using data!

**Data Description** — Show Map | Google Earth

*Citation:* Volostnykh, BV (1979): (Table 1) Weather conditions of the Western Sargasso Sea in October-November 1977. doi:10.1594/PANGAEA.755351,
*In Supplement to:* **Volostnykh, Boris V (1979):** Forms of phosphorus in the surface microlayer of the Western Sargasso Sea. *Oceanology*, **19(1)**, 44-46

*Project(s):* **Archive of Ocean Data** (ARCOD)

*Coverage:* *Median Latitude:* 29.128571 * *Median Longitude:* -70.461905 * *South-bound Latitude:* 28.016667 * *West-bound Longitude:* -71.383333 * *North-bound Latitude:* 29.983333 * *East-bound Longitude:* -69.283333

*Event(s):*
**VITYAZ7729** * *Latitude:* 28.333333 * *Longitude:* -71.300000 * *Date/Time Start:* 1981-11-08T08:45:00 * *Date/Time End:* 1981-11-20T11:30:00 * *Location:* Sargasso Sea * *Campaign:* VITYAZ * *Basis:* Vityaz * *Device:* Buoy
**VITYAZ7731** * *Latitude:* 29.633333 * *Longitude:* -71.383333 * *Date/Time Start:* 1981-11-05T08:15:00 * *Date/Time End:* 1981-11-21T03:10:00 * *Location:* Sargasso Sea * *Campaign:* VITYAZ * *Basis:* Vityaz * *Device:* Buoy
**VITYAZ7732** * *Latitude:* 29.983333 * *Longitude:* -70.650000 * *Date/Time Start:* 1981-11-04T13:14:00 * *Date/Time End:* 1981-11-21T09:52:00 * *Location:* Sargasso Sea * *Campaign:* VITYAZ * *Basis:* Vityaz * *Device:* Buoy


*Parameter(s):*

| # | Name | Short Name | Unit | Principal Investigator | Method | Comment |
|---|---|---|---|---|---|---|
| 1 | Event label | | | | | |
| 2 | LATITUDE | | | | | |
| 3 | LONGITUDE | | | | | |
| 4 | Date/Time of event | | | | | |
| 5 | Sample code/label | Label | | Volostnykh, Boris V | | |
| 6 | Wind direction | dd | deg | Volostnykh, Boris V | | |
| 7 | Wind speed | ff | m/s | Volostnykh, Boris V | | |
| 8 | Pressure, atmospheric | PPPP | hPa | Volostnykh, Boris V | | |
| 9 | Temperature, water | Temp | °C | Volostnykh, Boris V | | |
| 10 | Temperature, air | TTT | °C | Volostnykh, Boris V | | |
| 11 | Wave height | Wave height | m | Volostnykh, Boris V | | wind waves |
| 12 | State of the sea description | State sea descr | | Volostnykh, Boris V | | wind waves |
| 13 | Wave height | Wave height | m | Volostnykh, Boris V | | swell |
| 14 | State of the sea description | State sea descr | | Volostnykh, Boris V | | swell |

**http://dx.doi.org/10.1594/PANGAEA.755351?format=textfile**

*License:* Creative Commons Attribution 3.0 Unported

*Size:* 70 data points

**Download Data**

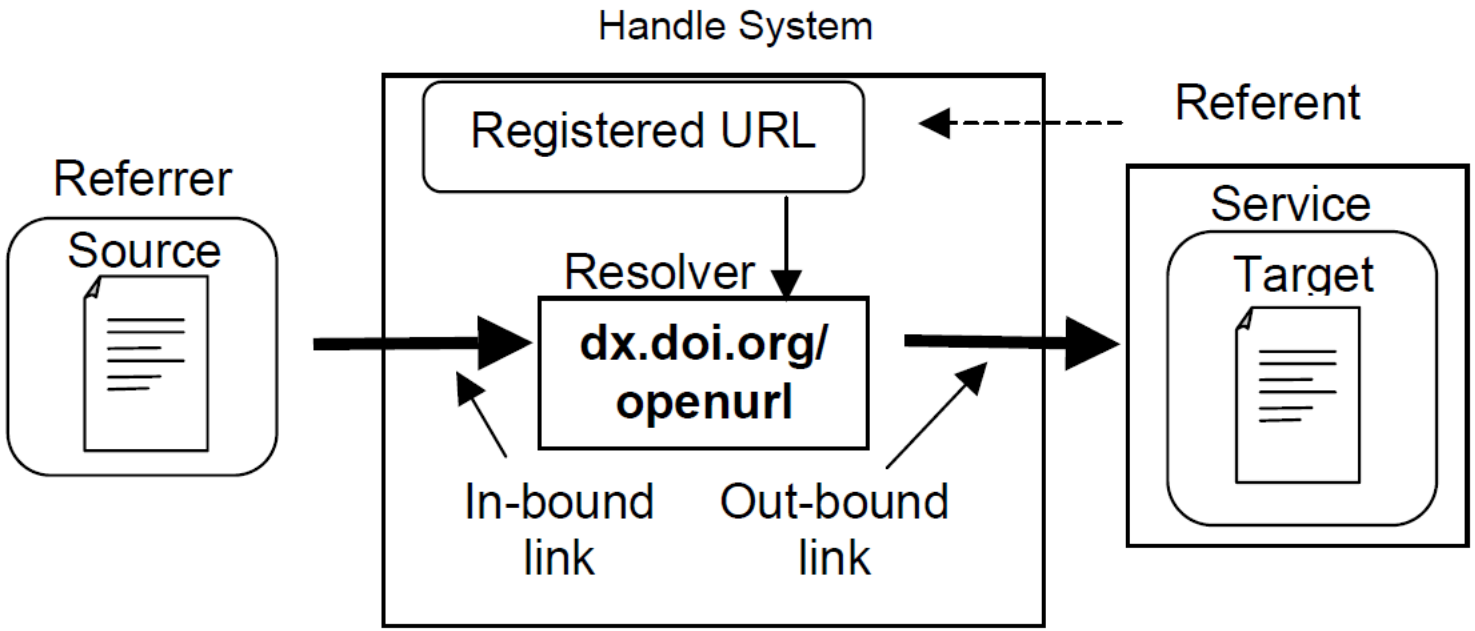
Download dataset as tab-delimited text (*use the following character encoding:* ISO-8859-1: ISO Western (PANGAEA default))

View dataset as HTML

# OpenURL

- rfr_dat - The referrer's parameter payload
  - Name value pairs
- rft_dat - The referent's parameter payload
  - Name value pairs

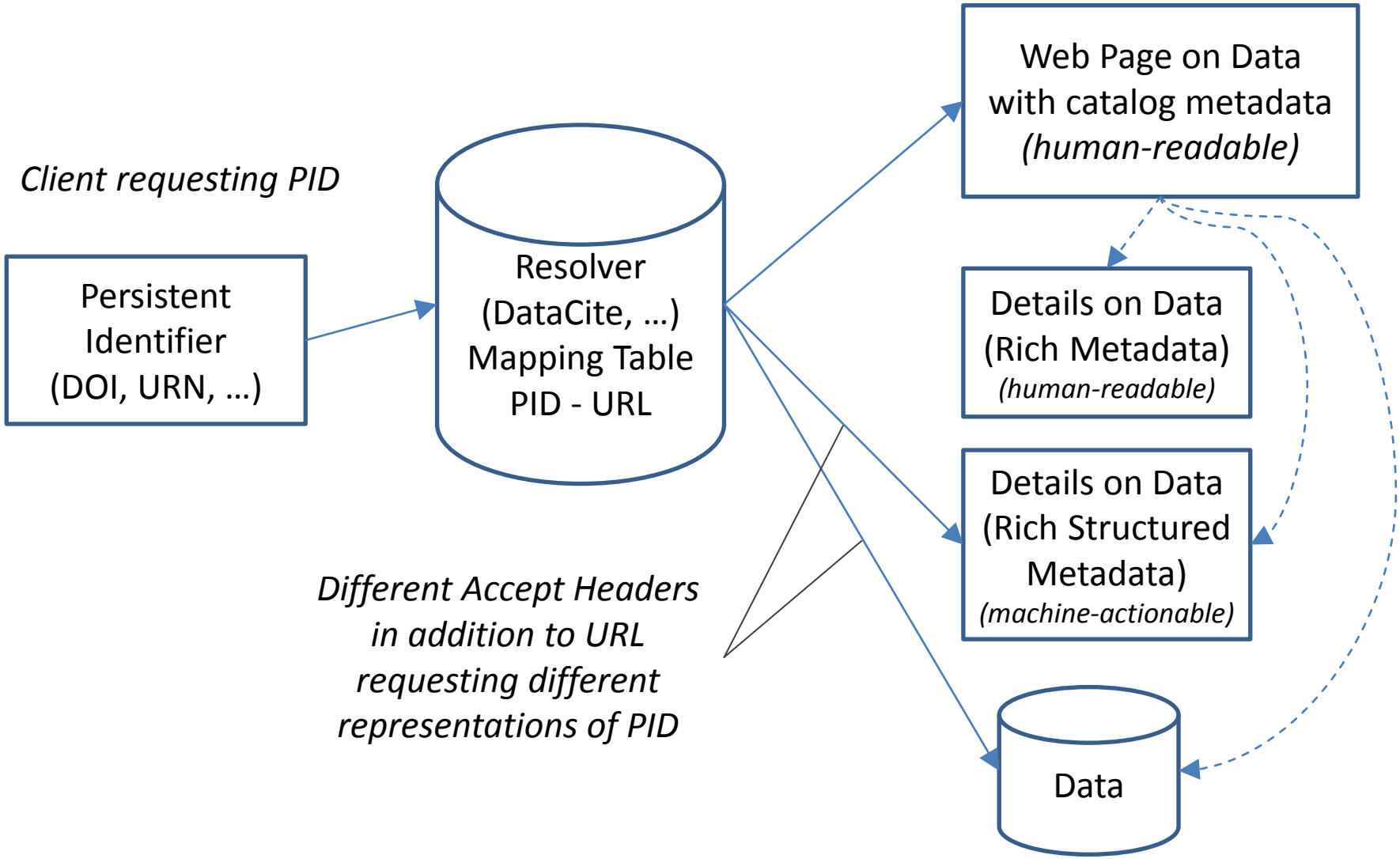
Reference: [Parameter Passing Via The DOI Proxy]



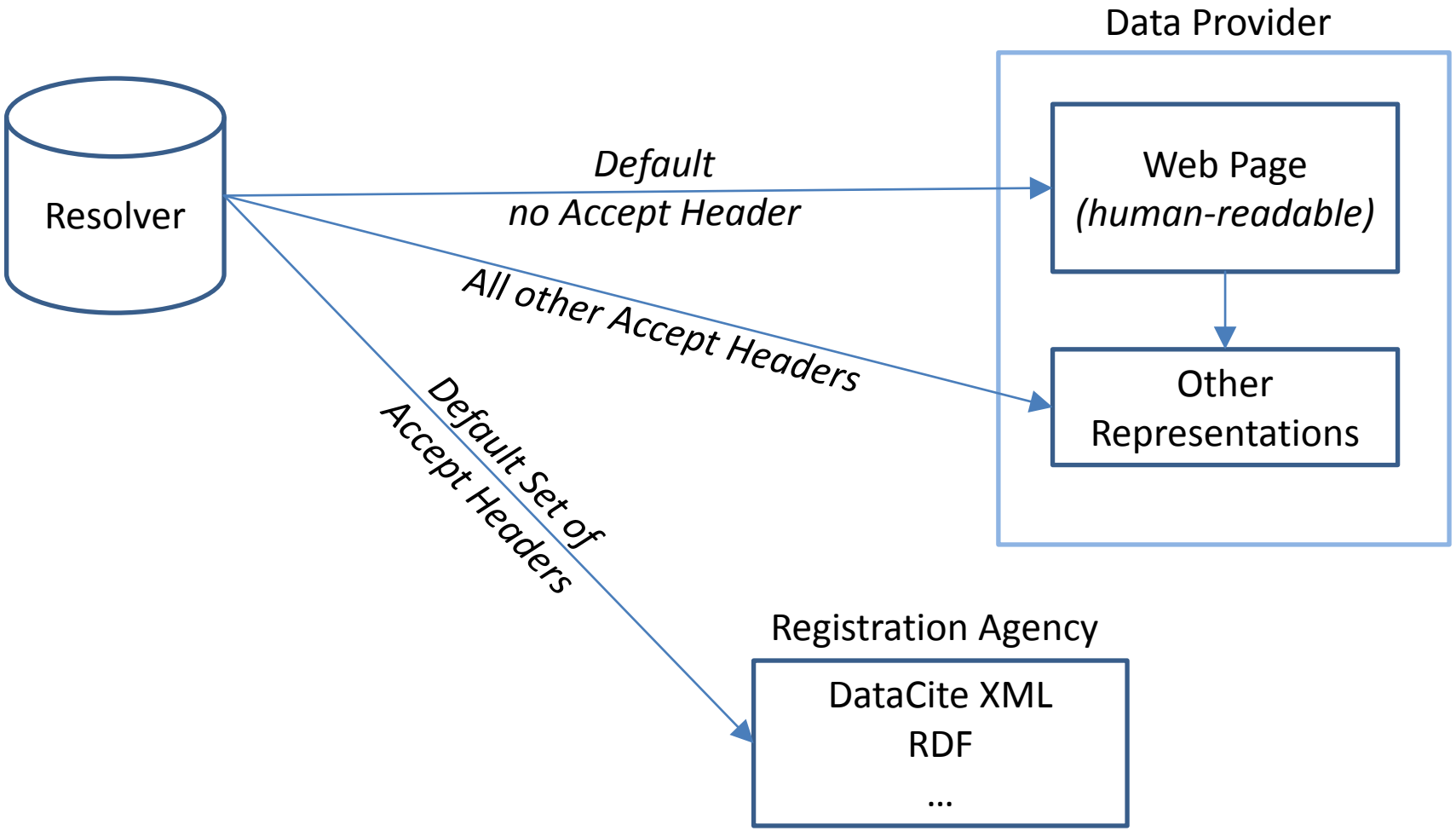
# HTTP Content Negotiation

- HTTP Content Negotiation is a method for HTTP clients to request different representations of an Internet resource
- Server-driven
  - Clients specify the desired media types (MIME type). Server responds according to request if possible, otherwise with a default response.
- Client-driven
  - Server sends the client a list of available representations of the requested resource and the client application selects the one to view.
- Different representations are not accessible by URLs
  - Not intendend for Web browser. Different solution required.

# Content Negotation - Based on Solution of CrossRef/DataCite



# Content Negotation
# Distinction of HTTP Accept Headers



# Content Negotiation Example

**http://dx.doi.org/10.5524/100005**

Response: Landing page

Redirected to data.datacite.org

```
http://dx.doi.org/10.5524/100005
Accept: application/x-datacite+text
```

Response: Text with citation information

Li, J; Zhang, G; Lambert, D; Wang, J; (2011): Genomic data from the Emperor penguin (Aptenodytes forsteri); GigaScience. http://dx.doi.org/10.5524/100005

Example DOI from DataCite Content Service (data.datacite.org)

# Content Negotiation Example (continued I)

```
http://dx.doi.org/10.5524/100005
Accept: application/x-datacite+xml
```

Response: DataCite XML

```
<resource …>
  <identifier identifierType="DOI">10.5524/100005</identifier>
  <creators>
    <creator>
      <creatorName>Li, J</creatorName>
    </creator>
    …
   <titles>
    <title>Genomic data from the Emperor penguin (Aptenodytes forsteri)</title>
  </titles>
  <publisher>GigaScience</publisher>
  <publicationYear>2011</publicationYear>
</resource>
```

# HTTP Accept Header

Accept: MIME type [; extension(s)]

Extensions are name/value pairs

- Extensions can be used for fine tuning.
- Extensions should come from a controlled vocabulary. No standard existent.

# MIME Media Types

- MIME - Multipurpose Internet Mail Extensions
  - extends the format of email to support such as non-text attachments
- MIME type could be understood as content type and/or a file format
- Examples
  - Web page: **`text/html; charset=UTF-8`**
  - PDF file: **`application/pdf`**
  - Binary data : **`application/octet-stream`**
  - Vendor-specific: **`application/vnd.`*`string`***
  - Non-standard: **`application/x-`*`string`***

# MIME Types - Limitations

- IANA (Internet Assigned Numbers Authority) maintains a registry
  - Important types are missing like CSV, statistical packages
- Apache Group maintains another list
  - Includes for example CSV, but nothing for statistical packages.
- Additional information is often required to process a file appropriately.
  - Example: Delimiter in CSV files.
- File extensions are a related approach, but can be ambigious (like ".rdf")

# HTML Links

```
http://data.datacite.org/ ↵
      application/x-datacite+xml ↵
      /10.5524/100005
```

Response: DataCite XML

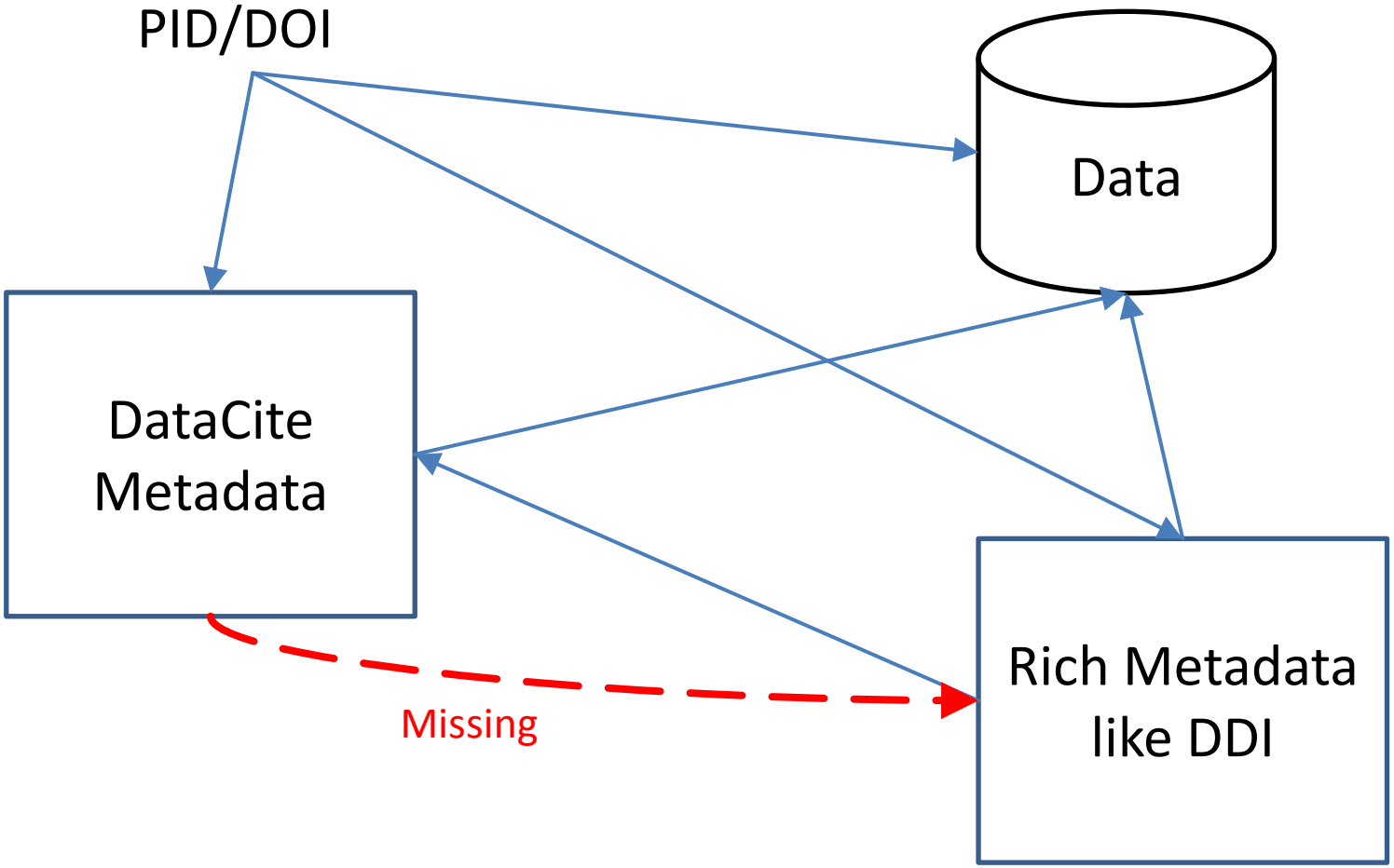
- Accessible by Web browsers. Workaround for this purpose.

- Purpose is just to provide a link. The form doesn‘t seem to conform REST principles. The core part of the URL of the resource changed.

  Possible alternative:

```
http://data.datacite.org/10.5524/100005? ↵
      mimetype=application/x-datacite+xml
```

# Relationship of DataCite Metadata / Rich Metadata / Data



# Conclusion I

- Techniques exist for a machine-actionable processing chain to access PID-related representations.
- Content negotiation seems to be a flexible approach.
  - Rules for redirection should be clarified between registration agency and data provider
  - MIME types are sometimes not sufficient. Values for Accept header extensions should come from controlled vocabularies
- Relationship from DataCite metadata to domain-specific metadata schemes need to be clarified.
  - Is there a specific property missing like:

    **RichMetadataIdentifier**

    richMetadataIdentifierName
    richMetadataIdentiferScheme

# Conclusion II

- Rich metadata seems to be important
  - What to do if this kind of metadata is not available?
- There seems to be a need for best practices for data providers such as
  - How to enable machine-actionable processing of metadata/data

# Thank you

joachim.wackerow@gesis.org