LTDP Long Term Data Preservation

Persistent Identifiers

Data Stewardship Interest Group

WGISS-37 Meeting

Cocoa Beach (Florida-US) - April 14-18, 2014

Overview

1. Introduction to Persistent Identifiers in Earth Sciences and Earth Observation
2. Needs, opportunities, challenges, considerations, systems
3. Implementation options
4. Next steps

Persistent Identifiers: generic need

- **Interoperability**: Persistent Identifiers (PI) are key for *interoperability* within a *community* and *among communities*.
- **Knowledge**: Persistent identification of digital objects (e.g. articles, datasets, images, stream of data) and non-digital objects (namely real-world entities, like authors, institutions but also teams, geographic locations and so on) is becoming a *crucial* issue for the whole information society.
- **Data Retrieval**: The functionality to *unambiguously locate* and *access* digital resources, associate them with the related authors and other relevant entities (e.g. institutions, research groups, projects) is becoming *essential* to allow the citation and retrieval of cultural and intellectual resources.
- **Preservation**: The rapid *increase* of digital assets, especially in the context of e-science, has made this dependency even stronger, making clear that *digital identifiers are crucial for preserving*, managing, accessing and re-using huge amounts of data over time.
- **Resource Discovery**: The implementation of a system for persistent identification of digital and non-digital objects is becoming a crucial *prerequisite* for sustained and reliable resource *discovery*, *citation and re-use*.
- **Unsuitable URLs**: (which have been adopted from the birth of the Web to identify and reference network resources) can not be considered a reliable approach to address the long term identification and access of digital resources due to the fact that *URLs serve the combined purpose of identifying a resource and describing its location*:
  - If the resource is moved to another location, the previous URL is no longer useful to access the resource.
  - Use of persistent identifiers has become *the most popular solution to preserve access to a digital resource regardless of its location*, by associating the persistent identifier with the correct current location, when the resource is moved.

Persistent Identifiers: ES & EO objectives & needs

Objective:

1. Assess if and how persistent identifiers can be introduced in Earth Sciences and to Earth Observation mission data
2. Persistent Identification: provide a globally unique, unambiguous, and permanent identification of a digital object for locating and accessing it for a long time

Needs:

1. To improve discoverability and accessibility
2. To enable users to retrieve objects without knowing their location
3. To enable repositories to change the location of objects internally
4. To enable repositories to share objects with other services where appropriate
5. To enable researchers to cite objects consistently over time

Persistent Identifiers: definition & requirements

Definition:

- Persistent identifiers (PIs) are simply identifiers that allow us to refer to a digital object, *location independent;*
- A Persistent Identifier is an identifier that is *effectively* and *permanently* assigned to an object.

PI system requirements:

- *Global uniqueness:* We consider the identifier a label that is associated with an object in a certain context. "Context" is intended as both the kind of standard used for the name syntax and the identification of the authority (sub-namespace) that assigns this label.
- *Persistence:* Persistence refers to the *permanent lifetime of an identifier*. It is not possible to reassign the PI to other resources or to delete it. That is, the PI will be globally unique forever
- *Resolvability:* refers to the possibility of retrieving a resource only if it is published.
- *Reliability*: the PI infrastructure must always *be active* (service redundancy, back-up deposit services, etc.) and the *register updated* (through automatic systems).
- *Authority*: is who *assign*, *manage* and *resolve* the identifiers.
- *Flexibility*: An identifier system will be more effective if it is *able to accommodate the special requirements of different types of material or collections*.
- *Interoperability:* this aspect is fundamental for guaranteeing the possibility of diffusing and accessing science digital objects.

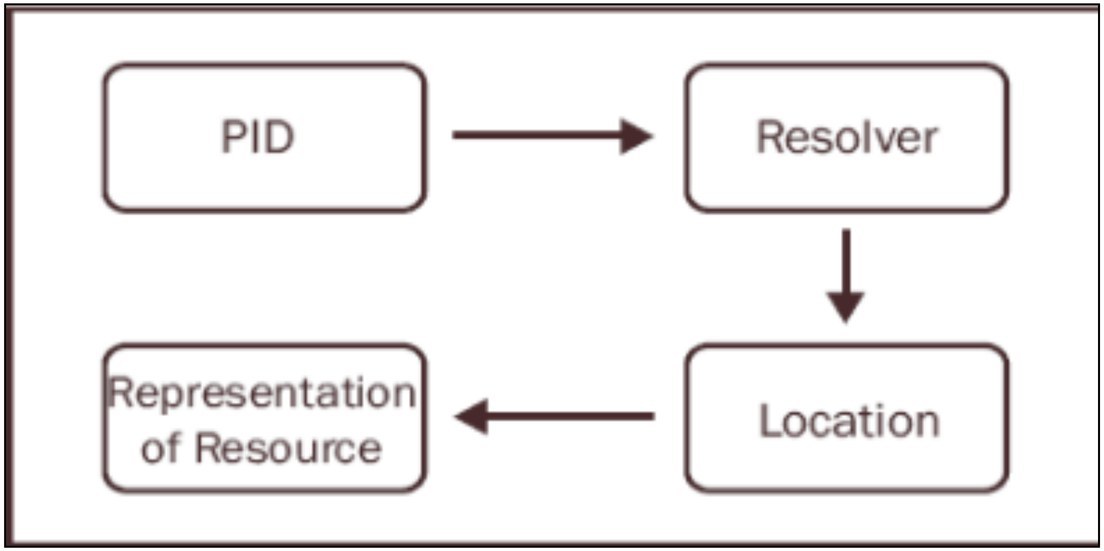
Persistent Identifiers: ES & EO benefits

Benefits for data holders

1. Increase data visibility
2. Show data use
3. Increase data use
4. Credibility / value of your data holdings

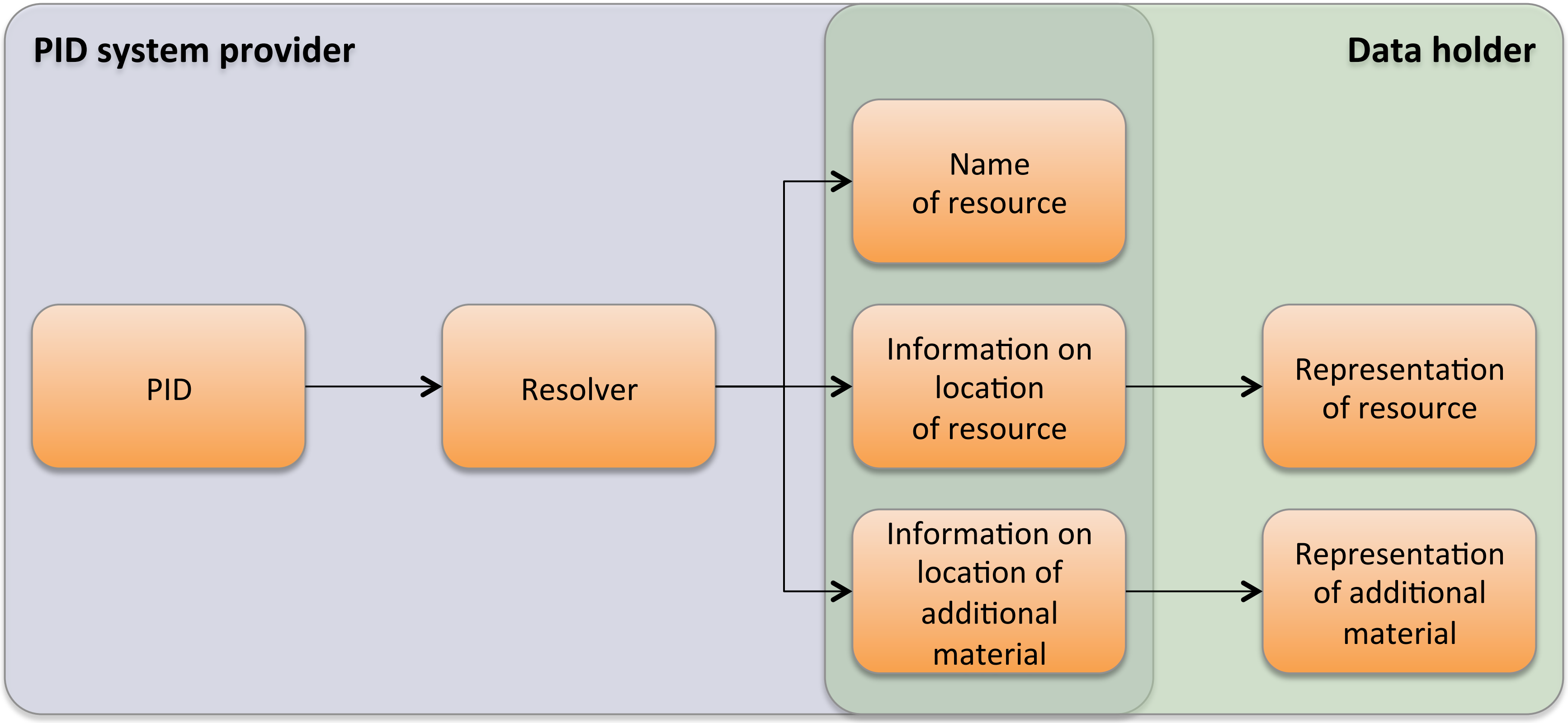
Persistent Identifiers: resolver

Almost all PIs System use a Resolver to retrieve the physical location



Example scenario: When a user uses an identifier to request a digital object, the location of it may be retrieved by a system to resolve locations from identifiers even when their location changes.

Persistent Identifiers - Shared Responsibility

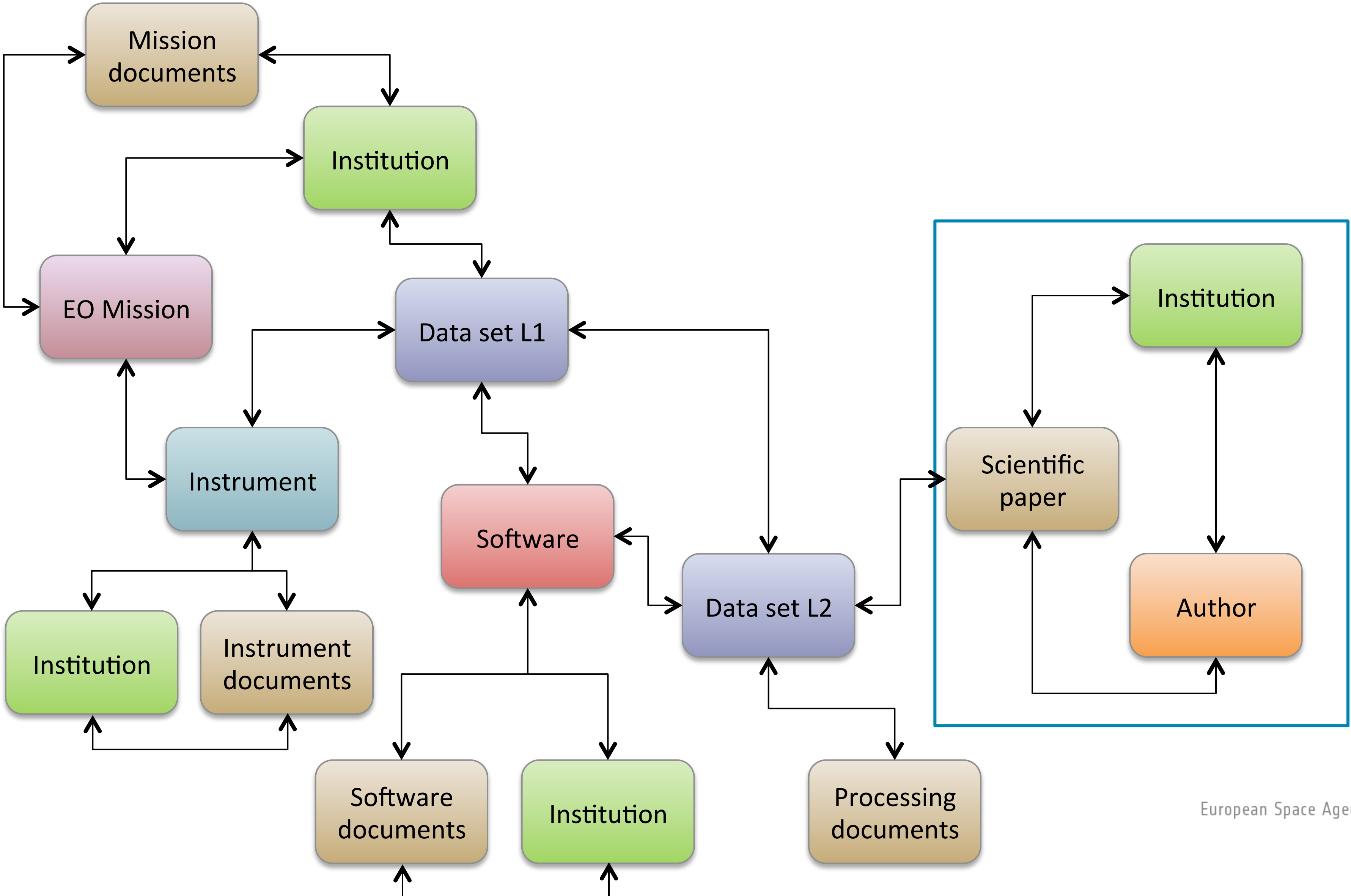

PID system provider
Data holder
PID
Resolver
Name of resource
Information on location of resource
Information on location of additional material
Representation of resource
Representation of additional material

Adapted from: http://www.paradigm.ac.uk/workbook/metadata/pids.html

Persistent Identifiers in EO Challenges and Considerations

- Long-term commitment → careful planning
- Access to the resource must be maintained
- Resource may move, be renamed, or be removed
- Resource may change, be updated → versioning capability
- What 'digital object' to assign PIDs to e.g. in Earth observation
  - Granularity - mission, instrument, product type ('collection'), scene, subset
  - Hierarchical PID system → large amount of PIDs
  - To which processing level? → raw, L0, L1, L2, …
  - To AIPs or DIP? → DIP is used but may get deleted after delivery
  - To data only or to all preserved information (LTDP PDSC) → PID dependency network
- Interoperability with other EO data providers

Possible Network of Relations of Persistent Identifiers in Earth Observation



European Space Agency

Examples of Persistent Identifier Systems

1. **ARK (Archival Resource Key)**
2. **DOI (Digital Object Identifier)**
3. LSID (Life Science Identifiers)
4. PURL (Persistent Uniform Resource Locator)
5. URI (Uniform Resource Identifier)
6. EPIC (European Persistent Identifier Consortium)
7. IGSN (International Geo Sample Number)
8. nbn-urn (National Bibliographic Numbers - Uniform Resource Name)

ARK: Archival Resource Key

esa

1. **An ARK ID is a URL**: http://library.manchester.ac.uk/ark:/98765/archive/object35
2. *Local* name resolver points browser to the object‘s current location
3. Object must have metadata, and a persistence statement
4. Developed & supported by University of California library
5. Data host has full control over the ID naming system
6. Can be hierarchical, versioning is supported
7. Advantages:
   - ✓no commercial interest
   - ✓low technical requirements
   - ✓under development, so we could influence the future path
8. Disadvantages:
   - ✓not widely used, not sure how popular it may become
   - ✓no financial resources to support the system
   - ✓higher potential that the system will disappear in the long run

DOI: Digital Object Identifier

1. **A DOI is a number inserted in the URL**:

   http://dx.doi.org/10.5067/MEASURES/DMSP-F11/SSMI/DATA303

2. *Central* name resolver points browser to the object's current location
3. Object must have metadata
4. Developed & supported by the DOI Foundation
5. Data host has full control over the ID naming system
6. Can be hierarchical, versioning is not supported
7. Advantages:
   - ✓free DOIs to public data providers
   - ✓most DOI users pay, so financial support for the whole DOI system
   - ✓widely used, becoming the global standard
   - ✓journals are citing data sources using DOI
   - ✓lower potential that the system will disappear in the long run
8. Disadvantages:
   - ✓no 100% guarantee that DOIs will remain free
   - ✓lower visibility, because the DOI uses doi.org in the URL

PID Implementation Examples in Earth Sciences and Earth Observation

- ARK implemented by:
  - ✓ CNES discovery tool
  - ✓ French National Library

- DOI implemented by:
  - ✓ UK Natural Environment Data Centre (NERC)
  - ✓ EUMETSAT Climate Centre
  - ✓ World Data Centre for Remote Sensing of the Atmosphere (WDC-RSAT) @ DLR, Germany
  - ✓ Parts of NOAA and NASA
  - ✓ Australian National Data Service

Observations

1. Lack of a standard and globally recognized solution
2. DOIs fulfill most criteria and seem to be the preferred PID system used in Earth Sciences and Earth observation
3. DOI is free of charge for any research group that has more than 50% government funding - at least if requested through Technical Information Library (TIB), Hannover, Germany
4. ARK solution is free
5. Harmonized guidelines for PID implementation - hierarchy, granularity, versioning, etc. - are more important than using a common PID system

LTDP WG Activities on PIDs

1. Preliminary Analysis of Persistent Identifiers already done
2. Comparative analysis of existing PIDs (e.g. DOI, etc) → DOI preferred

To be done:

1. Definition of approach and concept for use of PIDs in EO
2. Trade-off and cost/benefit analysis of implementation for the different solutions in EO
3. Proof of concept / pilot implementation
4. Recommendation

DSIG involvement

Technical Approaches for harmonization

1. Use same PID in all agencies
2. PIDs Translators between different PIDs systems
3. Harmonized Best Practice for PID implementation (hierarchy, granularity, versioning, etc.). Implementation left to each agency/organization.

Thank you for your attention !!!

Comments ? Questions ?

Persistent Identifiers: DOI (1)

The DOI system consists of four principal components:

- ✓ A naming syntax.
- ✓ A resolution service, based on the *Handle System*.
- ✓ A data model, which includes structured metadata based on a data dictionary and a framework for using this.
- ✓ Policies and procedures for implementing DOI names in a social infrastructure.

DOI syntax:

[Directory Code].[Registry code]/[Local Name] → (e.g. 10.7890/object786)

Directory Code The International DOI Foundation (IDF) is a Naming Authority under the Handle system; it has been allocated the number 10 as its unique identifier;

Registry Code is a unique number assigned by the IDF to an organisation that has been authorised to register DOI names - known as a Registration Agency (RA).

Local Name The local name suffix can be any alphanumeric string chosen by the *registering organisation*, which allows existing identification schemes to be incorporated into the DOI namespace.

The current location of each resource identified by a DOI is stored in the DOI system server, and any changes to this location must be registered there.

Persistent Identifiers: DOI (2)

Advantages

- The scheme is run by an *established* and *robust* organisation which is likely to be sustainable in the long-term.
- It has been adopted by libraries as well as commercial organisations.
- It provides an infrastructure for implementing a comprehensive digital identifier system.
- The possibility of establishing a 'Restricted' Application Profile means that the scheme could be used in a *non-public digital* repository environment or dark archive as well as an *open* environment.
- It is standards based and DOI metadata is created using XML, both of which maximise interoperability.

Disadvantages

- The annual subscription would be prohibitive.
- Whilst the DOI system offers a *sophisticated data model* which allows the creation of standardised metadata about digital resources and the grouping of resource-types into Application Profiles, these functions are probably *superfluous* to the needs of many curators looking after digital archives.
- DOI currently recommends using the scheme to identify only resources, parties and events associated with intellectual property transactions, whereas Paradigm has identified the need for a wider range of identifiers.

Persistent Identifiers: ARK (1)

The ARK scheme is founded on the principle that persistence is a matter of service, not syntax.

Each ARK is an actionable URL, which links users to:

- A digital object, although the scheme acknowledges that this kind of direct access may not be feasible (e.g. in dark archive).
- Metadata about that digital object.
- A commitment statement by the provider.

In order to assign ARKs, an institution must either become a Name Assigning Authority (NAA) under the scheme or be authorised to allocate names as a sub-authority of a NAA. Each NAA is associated with one or more *Name Mapping Authority Hostports* (NMAHs), which provide services.

ARKs work well with current protocols like HTTP and DNS, but they are designed to be protocol independent.

Persistent Identifiers: ARK (2)

ARK syntax:

[Protocol]/[NMAH/]ark:/[NAAN]/[Name]/[Qualifier] → (e.g. http://library.manchester.ac.uk/ark:/98765/archive/object35)

Protocol This label does not form part of the ARK identifier, but *indicates the protocol* which is being followed (e.g. http://).

NMAH This part of the string identifies the relevant NMAH or *provider* of services

ark:/ This prefix indicates where the actual ARK identifier begins.

NAAN stands for Name Assigning Authority Number: each NAA is assigned a 5 or 9 digit decimal number as a *unique identifier*. This element of the ARK string is *mandatory* because it unequivocally identifies the organisation which assigned the persistent name of the digital object.

Name is a mandatory element of the identifier and is assigned by the NAA. It should be *unique* within the NAA (ensuring its uniqueness within the system as a whole). *The NAAN and the Name taken together form the immutable persistent identifier for the object.*

Qualifier *optional* component of the ARK (e.g. identifying subcomponents or variants of a digital object).

Persistent Identifiers: ARK (3)

Advantages

- The scheme is *standards based and protocol/technology independent*.
- ARKs can be used to *identify different types of entity*, e.g. they could be used to identify agents and events as well as digital archival objects and metadata records.
- ARKs can be used in both a *closed environment* like a dark archive or an *open* publicly-accessible environment.
- The ARK system makes explicit the importance of organisational commitment to a persistent identifier scheme and writes a requirement for this into the scheme itself.
- It is maintained by a leading institution in the field of digital preservation and *has no commercially motivated background (like DOI)*.
- The model for participating in the ARK scheme is more *flexible* than some of the other PID schemes: if one institution acts as both NMAH and NAA, it is able to have complete control over its own identification scheme.
- The technical requirements for participation are relatively low: currently a normal web server using the DNS.
- Because the scheme is *still under development*, institutions which choose to participate now can feed into and shape this development.

Disadvantages

- The ARK scheme was *so recently* established and it is difficult to know at this stage *how popular* and long-lived it might be.
- Some elements of the scheme are probably *superfluous* to the requirements of digital archives.
- Most institutions are *moving towards encoding metadata in XML*, which is intended to be reasonably human-legible and facilitates the sharing of data across different information systems.