TRACKING DATASET TRANSFORMATIONS WITH HAPPI TOOLKIT

LUIGI BRIGUGLIO - BARI, NOVEMBER 11 2015
Presentation Topics

• Premise: where everything starts
• Digital Preservation: overview
• Tracking dataset transformations: datamodel
• HAPPI Toolkit: implementation
• Practice on HAPPI Toolkit @ EGI FedCloud
• Q&A
Premise: where everything starts

The **HAPPI Toolkit** is part of the **Data Preservation e-Infrastructure** produced by the **SCIDIP-ES project** [http://www.scidip-es.eu]

- This component, released with open source license (**Apache License v2.0**) and available on **SourceForge** [http://goo.gl/yWPBkV], is an implementation of an **authenticity model** defined by the collaboration of the **APARSEN** and **SCIDIP-ES** projects.

- This model describes how to **trace and document transformations** on any digital object during the whole life cycle, and it is based on **Open Provenance Model** and **PREMIS**. These de-facto standards improve interoperability among different **digital archives** and **communities**.

- Description of transformations on digital object is part of “**preservation metadata**” (a.k.a. Preservation Description Information) includes provenance, reference and integrity information, according to the **Open Archival Information System (OAIS)**, standard ISO:14721:2012.

---

**SCIDIP-ES**

*SCIENCE DATA INFRASTRUCTURE FOR PRESERVATION – EARTH SCIENCE*

**APA RSEN**

*Alliance Permanent Access to the Records of Science in Europe Network*

**#traceability**

**#OAIS**
Premise: where everything starts

ARCHIVE SETUP

DATA ACCESS

ARCHIVE EVOLUTION

Long-Term Digital Preservation Infrastructure

EGI CF2015 – Tracking Dataset Transformations with HAPPI Toolkit
Premise: where everything starts

ICT

Research

Earth Science Community

EGI CF2015 – Tracking Dataset Transformations with HAPPI Toolkit
Premise: where everything starts

• **APARSEN** proposed a methodology for the management of the authenticity of Digital Resources (DR):
  - **Formal authenticity model**: to represent the DR life cycle and the management of authenticity evidence
  - **Operational guidelines**: to guide the process of instantiating the model in a specific environment
  - **Case studies**: carried out to tune the methodology and test its effectiveness in a set of heterogeneous environments

• Cooperation among **APARSEN** (specifically La Sapienza University) and **SCIDIP-ES** (specifically Engineering) improved the model and produced its implementation: **HAPPI Toolkit**
Premise: where everything starts

HAPPI 1.5.0 instances run for validation in

SCIDIP-ES

SCIENCE DATA INFRASTRUCTURE FOR PRESERVATION - EARTH SCIENCE

EGI CF2015 – Tracking Dataset Transformations with HAPPI Toolkit
Presentation Topics

- Premise: where everything starts
- Digital Preservation: overview
- Tracking dataset transformations: datamodel
- HAPPI Toolkit: implementation
- Practice on HAPPI Toolkit @ EGI FedCloud
- Q&A
Digital Preservation: overview

- To promote standards for archiving (space) information, NASA has been involved in the CCSDS (Consultative Committee for Space Data Systems) and the ISO TC (Technical Committee) and SC (Sub-Committee):
  - TC 20: Aircraft and Space Vehicles
  - SC 13: Space Data and Information Transfer Systems

- Digital Preservation aims at ensuring digital information is accessible, understandable and usable over long time

- ISO:14721:2003 - Space data and information transfer systems - Open Archival Information System - Reference Model (OAIS RM)

- ISO:14721:2012: introduced further details on Preservation Description Information and Authenticity
Digital Preservation: overview

- OAIS provides an **Information Model** based on key concept of Information Package

  - **Content**
  - **Preservation Description**
  - **Information Package (xIP)**

- And a **Functional Model**

  - **Producers** → **SIP** (Ingestion (Submission)) → **AIP** (Archival Storage) → **DIP** (Access (Dissemination)) → **Consumers**
Digital Preservation: overview

- Information Package
- Descriptive Information
  - Metadata for retrieval
  - further described by
    - Content Information
      - Content to preserve
    - Preservation Description Information
      - Metadata for preservation
  - Reference
  - Fixity
  - Provenance
  - Context
  - Access Rights

EGI CF2015 – Tracking Dataset Transformations with HAPPI Toolkit
Presentation Topics

• Premise: where everything starts
• Digital Preservation: overview

**Tracking dataset transformations: datamodel**

• HAPPI Toolkit: implementation
• Practice on HAPPI Toolkit @ EGI FedCloud
• Q&A
• During its life cycle, data may undergo through many transformations (incl. changes of custody)
• Those transformations may affect the authenticity of data, for this reason it is important they are properly documented
• Evidences of transformations will be later used for authenticity assessment
Tracking dataset transformations: datamodel

- The datamodel of **HAPPI Toolkit** is based on the Authenticity Model defined by APARSEN and SCIDIP-ES
- Each Transformation is documented by a record, providing user with «evidence» of occurred events
Tracking dataset transformations: datamodel

- **HAPPI Toolkit** is a software component that manages part of preservation metadata defined in ISO:14721:2012, i.e. **OAIS** Preservation Description Information (PDI)

- This metadata is called EvidenceHistory and describes *evidences* for the transformations occurred on digital objects during their life cycle, that is **tracking transformations** on digital objects
Tracking dataset transformations: datamodel

Agent

Transformation

Representation

Intellectual Entity

EGI CF2015 – Tracking Dataset Transformations with HAPPI Toolkit

controlledBy

used

generatedBy

of

of
## Tracking dataset transformations: datamodel

<table>
<thead>
<tr>
<th>Intellectual Entity</th>
<th>Is a “coherent set of content that is described as a unit”, the goal of the preservation process being “to maintain usable versions of intellectual entities over time”.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Representation</td>
<td>Is a <strong>set of digital objects</strong> required to display, play, or otherwise make useable to a human a <strong>given version of an IE</strong>.</td>
</tr>
<tr>
<td>Transformation</td>
<td>Is a <strong>change</strong> that intervenes in conjunction with an event in the IE lifecycle, and produces a <strong>new representation of the IE</strong>, thus potentially <strong>affecting its authenticity</strong>.</td>
</tr>
<tr>
<td>Agent</td>
<td>Is the <strong>actor</strong> (human, machine, or software) associated with a given transformation of an IE, and <strong>who bears the responsibility</strong> of it.</td>
</tr>
</tbody>
</table>

EGI CF2015 – Tracking Dataset Transformations with HAPPI Toolkit
Tracking dataset transformations: datamodel

**Report**
- info
- Fixity
- SignificantProperties

**Agent**
- ID+info
- Type

**Transformation**
- ID+Info
- Software
- Type

**Representation**
- ID+info
- Format
- Type

EGI CF2015 – Tracking Dataset Transformations with HAPPI Toolkit
Tracking dataset transformations: datamodel

• To guarantee «interoperability» among communities and archives, data model has been based on:
  – **OPM**: Open Provenance Model – formalism for modelling life cycle of digital object as a provenance graph

  – **PREMIS**: Data Dictionary for Preservation Metadata – common dictionary in the preservation community for ensuring interoperability among repositories
Tracking dataset transformations: datamodel
Tracking dataset transformations: datamodel

- Some transformations change the intellectual entity and generate new one(s), e.g.
  - Extraction
  - Aggregation
Presentation Topics

• Premise: where everything starts
• Digital Preservation: overview
• Tracking dataset transformations: datamodel
• HAPPI Toolkit: implementation
• Practice on HAPPI Toolkit @ EGI FedCloud
• Q&A
HAPPI Toolkit: implementation

- **HAPPI** (Handling Authenticity Provenance and Persistent Identifiers)
  - Manage Intellectual Entity
  - Capture Evidence Record Documentation (OPM1.1 and PREMIS2.2)
  - Store Intellectual Entity, Evidence Record/History in a scalable database
  - Search/Browse
  - Import/Export
HAPPI Toolkit: implementation

- Archive Manager can add specific significant properties, for later supporting authenticity assessment.

- Reference is applied to Intellectual Entities and evidence items (i.e. Agent, Transformation, Representation):
  - Organisation – who assigns the reference
  - Type – type of reference (e.g. URI, DOI)
  - Value – value of reference

- Type of Transformations
  - AGGREGATION
  - CAPTURE
  - CHANGEOF Custody
  - EXTRACTION
  - INGESTION
  - MIGRATION
HAPPI Toolkit: implementation

**Register the Intellectual Entity**
- title
- creation date
- reference
- annotation

**Step 1**

**Gather information into Evidence Records**
- transformation
- who controls the transformation
- result/input of transformation
- report with annotation and specific properties

**Step 2**
HAPPI Toolkit: implementation

HAPPI-LOGIC-1.5.0

IEManager

EHManager

IntellectualEntityManager

- addIntellectualEntity(ie)
- getIntellectualEntity(label)
- getAllIntellectualEntities()
- getIntellectualEntitiesBy(from, to, keyword)

EvidenceHistoryManager

- addEvidenceRecord(er, eh)
- getEvidenceRecord(label, eh)
- getAllEvidenceRecords(eh)
- getEvidenceRecordHistory(label, eh)
- getLastEvidenceRecords(eh)
- importEvidenceHistory(eh, gxmlFile)
- exportEvidenceHistory(eh, gxmlFile)
HAPPI Toolkit: implementation

HAPPI-SERVER

HAPPI-LOGIC

Blueprints

OrientDB

GraphDB

Neo4j
HAPPI Toolkit: implementation

Browse the History of Data
Timeline mode

ENVISAT MERIS Level 2 Full Resolution (MER_FR__0P)
Was controlled by Luigi Brigugio
Was generated by Transformation from L0 to L1
Annotation: "Captured L1 dataset from ENVRI portal"
For further details see evidence report
Link to dataset from ESA

ENVISAT MERIS Level 1 Full Resolution (MER_FR__0P)
HAPPI Toolkit: implementation

Browse the History of Data

Graph mode
// obtain the IntellectualEntityManager
IntellectualEntityManager iemanager = ManagerFactory.getInstance().getIntellectualEntityManager();

// create the reference

/**
 * create the intellectual entity, that is composed by
 * reference, label, title, annotation and date of creation.
 */
IntellectualEntity ie1 = new IntellectualEntity(sampleRef, null,
"HAPPI Infographics",
"SCIDIP-ES HAPPI Infographics",
new Date());

// add the intellectual entity through iemanager
iemanager.addIntellectualEntity(ie1);
HAPPI Toolkit: implementation

```java
// obtain the EvidenceHistoryManager
EvidenceHistoryManager ehmanager =
    ManagerFactory.getInstance().getEvidenceHistoryManager();

// get the evidence history of the intellectual entity
EvidenceHistory eh1 = ie1.getEvidenceHistory();

/** create the first evidence record with sample data, by
 * using the buildRecord utility method.
 */
EvidenceRecord er1 = new EvidenceRecord();
er1 = buildSampleRecord("Luigi Briguglio", "capture","origin", "er1", null);

// add the evidence record to its history
eh1.addEvidenceRecord(er1);
ehmanager.addEvidenceRecord(er1, eh1);

// adding a second record to the history to the first one
EvidenceRecord er2 = new EvidenceRecord();
er2 = buildSampleRecord("Luigi Briguglio", "ingestion","submitted", "er2", er1);
eh1.addEvidenceRecord(er2);
ehmanager.addEvidenceRecord(er2, eh1);
```

http://sourceforge.net/p/digitalpreserve/code/HEAD/tree/SCIDIP-ES/software/toolkits/authenticity
Presentation Topics

• Premise: where everything starts
• Digital Preservation: overview
• Tracking dataset transformations: datamodel
• HAPPI Toolkit: implementation
• Practice on HAPPI Toolkit @ EGI FedCloud
• Q&A
Presentation Topics

• Premise: where everything starts
• Digital Preservation: overview
• Tracking dataset transformations: datamodel
• HAPPI Toolkit: implementation
• Practice on HAPPI Toolkit @ EGI FedCloud
• Q&A
References


• Thesis - Analisi Progettazione e Sviluppo di un Prototipo per la Gestione della Provenienza nel Processo di Conservazione Digitale, Tor Vergata Univ., October 2013


Next Step: Extending the model

- add relationships to other digital objects ➔ to document context
- document rights
- to better tracking evolution
Questions?