Digital Repositories:
Semantic Web
Linked Data
Knowledge Graphs

Una Década de Innovación Digital: Celebrando los 10 años del Repositorio Digital de la CEPAL, 2024

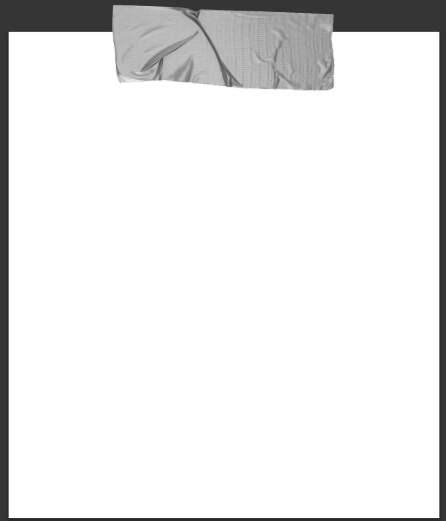
Claudio Gutierrez DCC, UChile & IMFD

(too) many keywords

Data, information, Knowledge, Repository, Database, Knowledge Base, Semantic Web, Linked Data, Knowledge Graph, …

Digital Repositories

(Collections of digital objects)



Digital repositories

• content is deposited in a repository, whether by the content creator, owner or third party

• the repository architecture manages content as well as metadata

• the repository offers a minimum set of basic services e.g. put, get, search, access control

• the repository must be sustainable and trusted, well-supported and well-managed (over time)

Heery, R & Anderson, S 2005, Digital repositories review. Joint Information Systems Committee

Open Access Repositories

Open access repositories can be distinguished by the following characteristics:

- the repository must provide open access to its content (unless there are legal constraints)
- the repository must provide open access to its metadata for harvesting

FAIR

The FAIR Data Principles highlight the need to embrace good practice by defining essential characteristics of data objects to ensure that data are reusable by humans and machines: they should be Findable, Accessible, Interoperable, and Reusable, i.e. FAIR. However, to make data FAIR whilst preserving them over time requires trustworthy digital repositories (TDRs) with sustainable governance and organizational frameworks, reliable infrastructure, and comprehensive policies supporting community-agreed practices.

Preservation requires trustworthy digital repositories (TDRs)

Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018 (2016).

OASIS

Open Archival Information System (OAIS) reference model provides recommendations on setting up archives delivering long-term preservation of and access to information (in particular, digital information) and creating preservation packages.

It offers a coherent and comprehensive framework of principles and terminology for the management of archival information systems.

However, conforming to the OAIS reference model does not guarantee trustworthiness.

Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). Recommended Practice CCSDS 650.0-M-2. *Consultative Committee for Space Data Systems*, (2012)

Certification

To assess and improve the quality of their professional practices, repositories rely on a range of international certification standards covering core, extended or formal level certification. These standards such as the CoreTrustSeal[6], DIN31644/NESTOR[7], and ISO16363[8] focus on four major assessment areas: organization, digital object management, technical infrastructure, and security risk management.

CoreTrustSeal. CoreTrustSeal Certified Repositories. *CoreTrustSeal*, https://www.coretrustseal.org/why-certification/certified-repositories/ (2020).

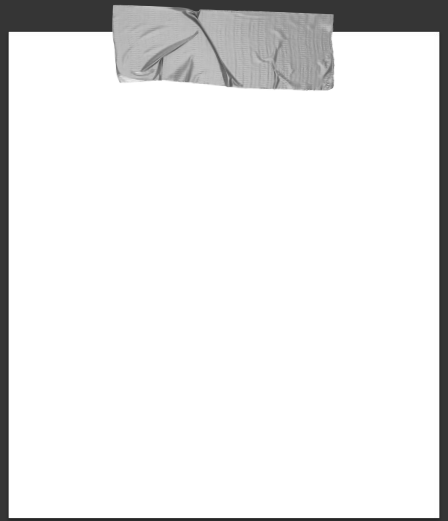
TRUST Principles

Principle	Guidance for repositories
Transparency	To be transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence.
Responsibility	To be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service.
User Focus	To ensure that the data management norms and expectations of target user communities are met.
Sustainability	To sustain services and preserve data holdings for the long-term.
Technology	To provide infrastructure and capabilities to support secure, persistent, and reliable services.

Lin, D., Crabtree, J., Dillo, I. *et al.* The TRUST Principles for digital repositories. *Sci Data* 7, 144 (2020).

Semantic Web

(Infrastructure for machine readable semantics)



Machine Readable Semantics

For a world at a scale beyond humans

Need to “collaborate” with machines on the semantic realm

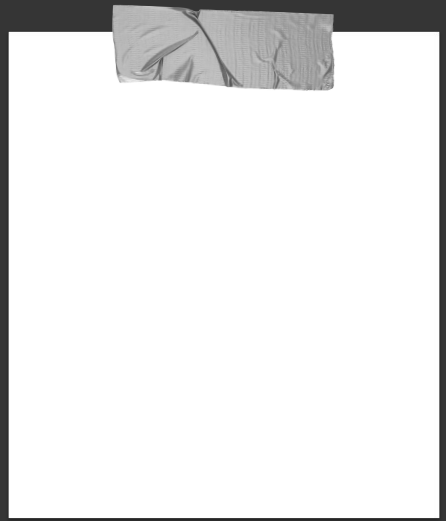
IDEA:

- Tag digital objects with formal metadata
- Sophisticated (logical) organization of metadata (vocabularies, taxonomies, ontologies)
- Develop machinery to retrieve digital objects

T. Berners-Lee, J. Hendler, O. Lassila. The Semantic Web: Scientific American, May 2001

Linked Data

(Semantic network of data over the Web)



Key idea:

Standardize and automate *semantic relationships* among digital objects

Take advantage of the Web infrastructure whose main idea is the (hyper) “link”

Create network of related digital objects over the Web

Develop machinery to organize, develop and query this network

Advantages

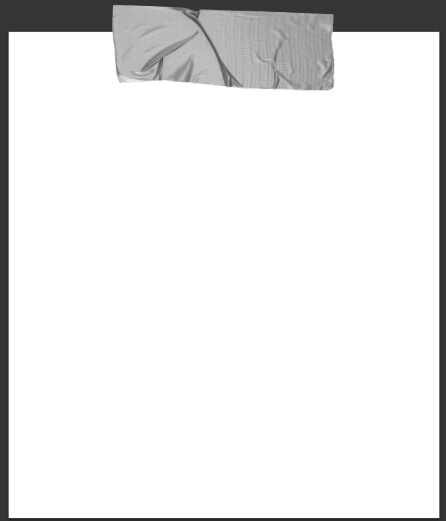
- Well known principles
- Universal architecture
- Excellent for open data
- Much better performance than traditional databases
- Interoperability through eg. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)

Limitation: Repositories that need partial access policies (e.g. enterprises, governments, personal, etc.)

Heath, T., Bizer, C. (2011). Principles of Linked Data.

Knowledge Graphs

(Semantic networks with 21st century digital technology)



Key idea: graphs

“relationship” is a more general notion than the (Web) link

Graphs (networks) are well understood mathematical objects

Can develop more flexible and diverse architectures (than SW and LD) that

- automate formal semantics (SW)
- link digital objects (LD)
- have diverse modes of governance (open, commons, etc)
- integrate technologies from DB, IR, SW, KR, LD, etc.

Knowledge graphs

***Core idea**:* using graphs to represent data, often enhanced with some way to explicitly represent knowledge. The result is most often used in application scenarios that involve integrating, managing and extracting value from diverse sources of data at large scale.

Knowledge Graphs: Systems that support networks of digital objects with automated semantics at different scales using diverse protocols and retrieval systems.

Advantages

- Well understood and friendly model
- Web still is the “universal” architecture
- Allow intermediate size - repositories
- Allows other types of governance and property rights (Community Governance, private organizations, etc.)
- Ample variety of access mechanisms
- Infrastructure for reasoning over the data
- Flexible architecture that use well known technologies (such as DB, KB, SW, LD, SN, IS, etc.)

Aidan Hogan et al. Knowledge Graphs. ACM Computing Surveys, 2021

some challenges

(as illustration)

- Diversity of digital objects
- Diversity of needs (storing, preservation, retrieval, access, etc.)
- Size beyond human scale
- Interoperability
- Evolution over time
- Property & privacy issues
- Learning & LLMs
- Models of funding
- etc.

Gracias por la atención

Claudio Gutierrez

cgutierr@dcc.uchile.cl