CERN

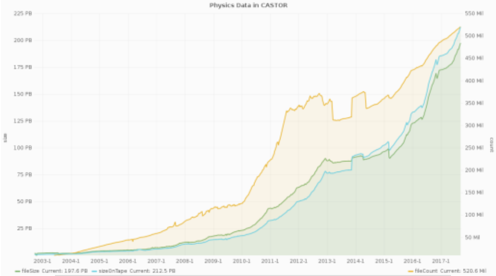
International Collaboration for **Data Preservation** and **Long Term Analysis** in High Energy Physics

Data Management and Access Policies: CERN, HEP (and beyond)

OECD-GSF Workshop on Open Data
October 2017

Jamie.Shiers@cern.ch

Outline

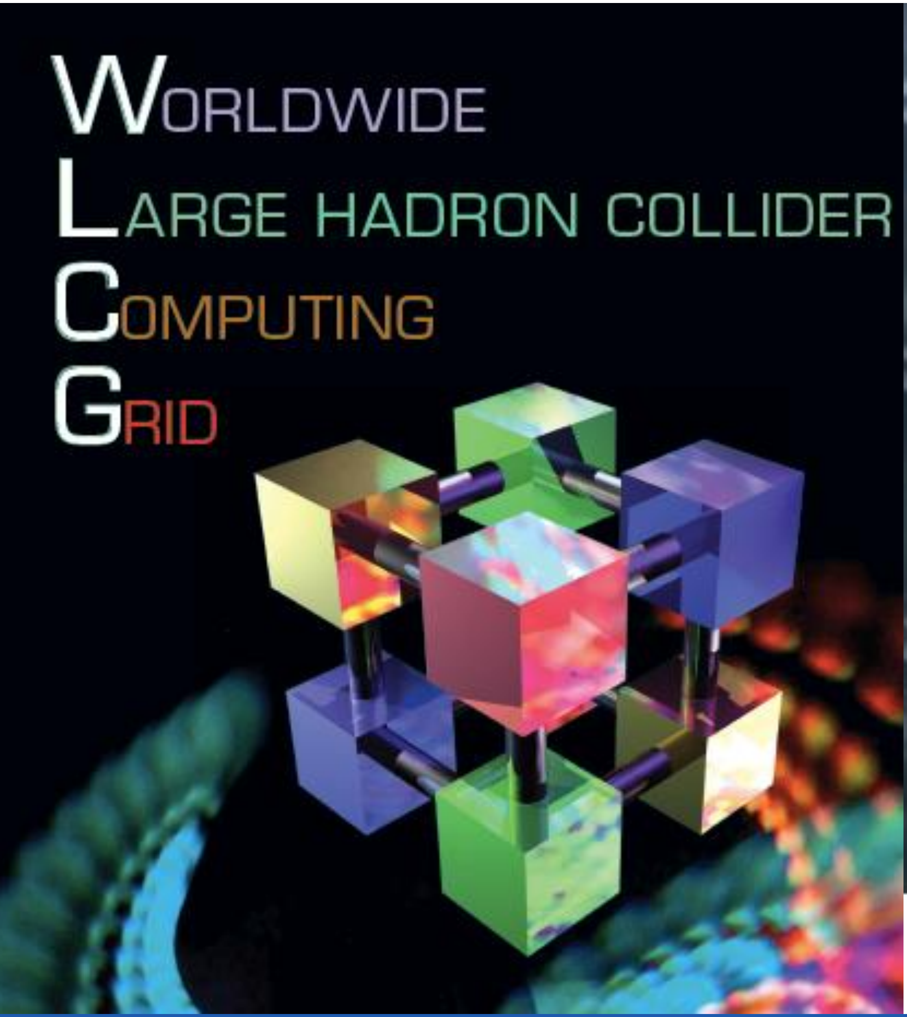


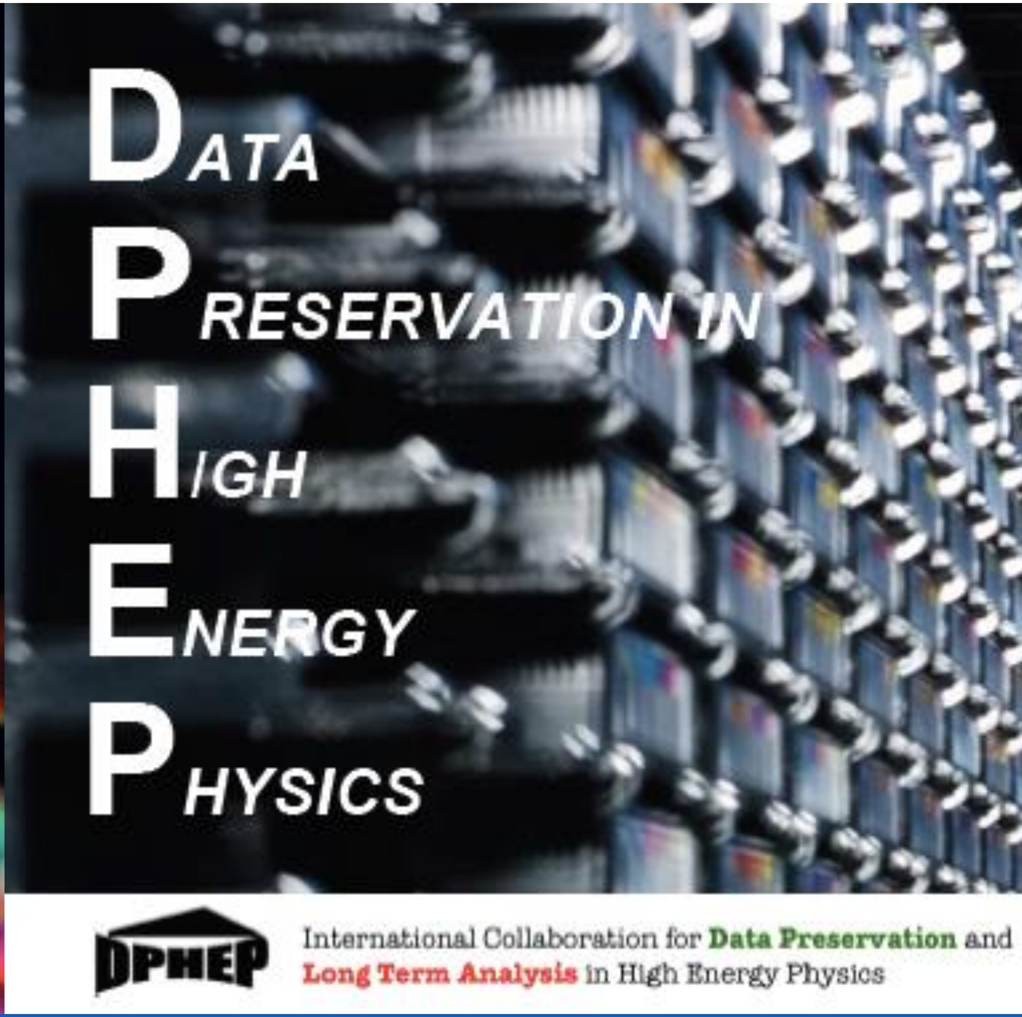
212 PB on tape

- ***What does Open Data (or FAIR principles) mean with regards to the massive amounts of data that are generated from experiments in particle physics or the planned SKA telescope?***

➢ **Focus on LHC, HL-LHC, plus also status at other HEP labs and possibilities for ESFRI / EIROForum collaboration(s)**

Data Management / Access Policies




DATA
PRESERVATION IN
HIGH
ENERGY
PHYSICS
DPHEP
International Collaboration for Data Preservation and
Long Term Analysis in High Energy Physics

You can’t share data, nor re-use it, unless you have preserved it!

FAIR Data Principles

Expert Group on turning FAIR into reality

TO BE **FINDABLE**:

- F1. (meta)data are assigned a **globally unique and eternally persistent identifier.**
- F2. data are described with rich metadata.
- F3. (meta)data are registered or indexed in a searchable resource.
- F4. metadata specify the data identifier.

TO BE **ACCESSIBLE**:

- A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
- A1.1 the protocol is open, free, and universally implementable.
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- A2 metadata are accessible, even when the data are no longer available.

TO BE **INTEROPERABLE**:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

TO BE **RE-USABLE**:

- R1. meta(data) have a plurality of accurate and relevant attributes.
- R1.1. (meta)data are released with a clear and accessible data usage license.
- R1.2. (meta)data are associated with their provenance.
- R1.3. (meta)data meet domain-relevant community standards.

From https://www.force11.org/group/fairgroup/fairprinciples

FAIR DMPs & TDRs

- *If we want to be able to* ***share data****, we need to store them in a* ***Trustworthy Digital Repository*** *(TDR).*
  - *Data created and used by scientists should be managed, curated, and archived in such a way to preserve the initial investment in collecting them.*
  - *Researchers must be certain that data held in archives remain useful and meaningful into the future.*
  - *Funding authorities increasingly require continued access to data produced by the projects they fund, and have made this an important element in* ***Data Management Plans*** *(DMPs).*
  - *Indeed, some funders now stipulate that the data they fund must be deposited in a trustworthy repository.*

Source: Ingrid Dillo, iPRES 2017 abstract (DANS and interim RDA SG)

Typical EU H2020 Call Text

- *Research Infrastructures, such as the ones on* ***the ESFRI roadmap*** *and others, are characterised by the* ***very significant data volumes*** *they generate and handle.*
- *These data are of interest to* ***thousands*** *of researchers across scientific disciplines and to other potential users via* ***Open Access*** *policies.*

➢ ***Effective data preservation and open access for immediate and future sharing and re-use are a fundamental component of today's research infrastructures.***

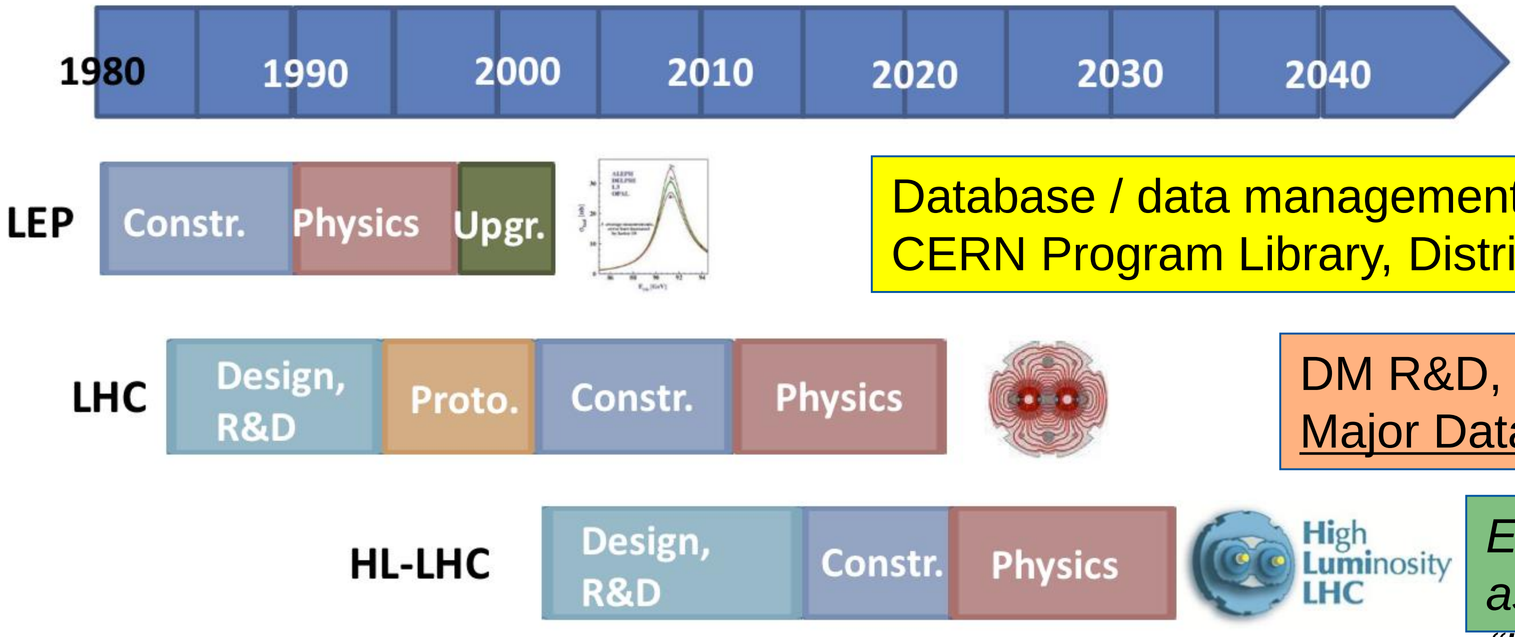
http://www.esfri.eu/roadmap-2016 HL-LHC is a Landmark project, as is SKA

CERN as a "TDR" (ISO 16363)

- We believe **certification** will allow us to ensure that best practices are implemented and followed up on in the long-term: "**written into fabric of organisation**"
- Scope: **Scientific Data** and CERN's **Digital Memory**
- **Timescale**: complete prior to 2019/2020 ESPP update
- Will also "ensure" adequate resources, staffing, training, succession plans etc.
- CERN can expect to exist until HL/HE LHC (2040/50)
- And beyond? FCC? Depends on physics…

Current ESPP (approved May 2013):
(i) data preservation […] should be maintained and further developed.

LEP / (HL-)LHC Timeline



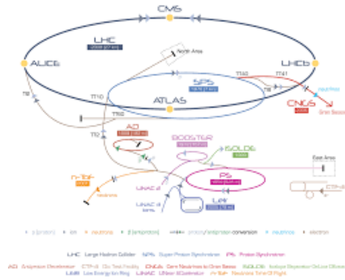
Database / data management support,
CERN Program Library, Distributed Computing

DM R&D, DBs, WLCG, EGI
Major Data Migrations(!)

ESFRI roadmap as “landmark project”

- Robust, stable services over **several decades**
- Data preservation and re-use over **similar periods**
- “Transparent” and supported **migrations**

Open Data at CERN



- The 4 main LHC experiments have approved **Open access** policies whereby (increasing) fractions of their data are made available after suitable “embargo periods”

➤ **These refer to “derived data” + documentation + s/w and environment**

  - The 3 main pillars of LTDP in HEP

❑ **But LHC data volume is already >200PB!**

  - Expected to reach ~10(-100)EB during HL-LHC!
  - We need to **preserve** all of this (but not all is **Open**)

See http://opendata.cern.ch - about (This portal; CMS; …)

LTDP: How do we measure progress / success?

- **Practice:** through Open Data releases
  - Can the data really be (re-)used by the Designated Community(ies)?
  - What are the support costs?
  - Is this sustainable?

- **Theory:** by applying state of the art "preservation principles"
  - Measured through ISO 16363 (self-) certification and associated policies and strategies
  - Participation in relevant working & interest groups

One, without the other, is probably not enough. The two together should provide a pretty robust measurement...

N.B. neither are one offs and need to be regularly repeated!

Data Preservation in High Energy Physics

http://dphep.org

- **LTDP in HEP includes: data, documentation, s/w + environment (and some commonality in services themselves)**

➢ **Open Access currently for LHC experiments – hard to apply this to past ones (who should one ask?)**

➢ **Plan to make this the default for future (CERN) experiments through Certification & DMPs**

H1
ZEUS
hermes
BELLE
CDF
DØ
ALICE
ATLAS
CMS
LHCb
DESY
KEK
Institute of High Energy Physics
Chinese Academy of Sciences
Fermilab
SLAC
NATIONAL ACCELERATOR LABORATORY
CERN
Jefferson Lab
Science & Technology Facilities Council
Max-Planck-Institut für Physik
INSPIRE
BROOKHAVEN
NATIONAL LABORATORY

2020 Vision for LTDP in HEP

- ***Long-term – e.g. FCC timescales*: disruptive change**

  - By 2020, all **archived data** – e.g. that described in DPHEP Blueprint, including LHC data – easily **findable**, fully **usable** by **designated communities** with clear (Open) access policies and possibilities to annotate further

  - Best practices, tools and services well run-in, fully documented and sustainable; built in common with **other disciplines**, based on standards

  - **DPHEP portal**, through which data / tools accessed
    - ➢ **"HEP FAIRport": Findable, Accessible, Interoperable, Re-usable**

➢ **Agree with Funding Agencies clear targets & metrics**

How Has FAIR evolved in 2017?

- Increasingly, FAIR has been taken to include not just data + meta-data but also software
- What started as "source code" preservation has now evolved to **"running s/w and its environment"**
  - **Much better IMHO**
- But there is still a lot to define / do
  - ➢ **How is the data Findable?**
    - **Navigation?** Search? Is there an API? …
  - **How to implement this in a scalable & sustainable way**
    - E.g. how many PID / DOI lookups per unit time, for how long is the service "guaranteed", … **"eternally?"**
  - **How to implement cross project / discipline searches?**

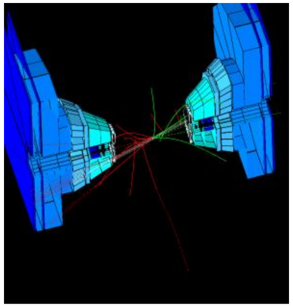
➢ **I have heard claims that people have been doing this for 20 – 100(!!!) years**

- ➢ **(These people clearly don't need any more project money)**

"As Open as possible; as closed as necessary"

~30 years of LEP – what does it tell us?

- Major migrations are **unavoidable** but hard to **foresee**!
- **Data** is not just "**bits**", but also **documentation, software + environment + "knowledge"**
  - "Collective knowledge" particularly hard to capture (remember)
    - Documentation "refreshed" after 20 years (1995) – now in Digital Library in PDF & PDF/A formats (was Postscript)
- Today's "**Big Data**" may become tomorrow's "**peanuts**"
  - **100TB** per LEP experiment: immensely challenging at the time; now "trivial" for both CPU and storage
  - With time, **hardware costs** tend to zero
    - O(CHF 1000) per experiment per year for archive storage
  - **Personnel costs** tend to O(1FTE) **>> CHF 1000!**
    - Perhaps as little now as 0.1 – 0.2 FTE per LEP experiment to keep data + s/w alive – no new analyses included

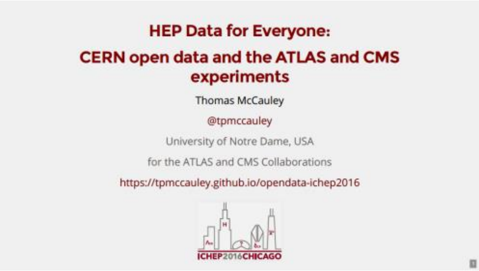


See DPHEP Workshop on "Full Costs of Curation", January 2014: https://indico.cern.ch/event/276820/

ODBMS migration – overview (300TB)

- **A triple migration!**
  - Data format and software conversion from Objectivity/DB to Oracle
  - Physical media migration from StorageTek 9940A to 9940B tapes
- **Took ~1 year to prepare; ~1 year to execute**
- **Could never have been achieved without extensive system, database and application support!**
- Two experiments – many software packages and data sets
  - **COMPASS** raw event data (300 TB)
    - Data taking continued after the migration, using the new Oracle software
  - **HARP** raw event data (30 TB), event collections and conditions data
    - Data taking stopped in 2002, no need to port event writing infrastructure
  - In both cases, the migration was during the "lifetime" of the experiment
  - System integration tests validating read-back from the new storage

Summary – Open Data



HEP Data for Everyone:

CERN open data and the ATLAS and CMS experiments

Thomas McCauley

@tpmccauley

University of Notre Dame, USA

for the ATLAS and CMS Collaborations

https://tpmccauley.github.io/opendata-ichep2016

ICHEP2016CHICAGO

- "Open Data" – today at multi x00TB – is a **reality** for the LHC experiments and will (hopefully) spread to all new CERN experiments (and beyond)

➢ **We see (BIG) benefits in making the data open: including ensuring the data is re-usable!**
  - For us as well as others (theorists, students etc.)

- **The additional costs are minimal (in comparison) – except for h/w resources which can be significant in the short-medium term**

❑ **We see clear opportunities for collaboration with related disciplines on this and wider DM aspects (EOSC and beyond)**

Possible Future Policy Work

- Understand how intra- and **inter-disciplinary** “FAIR DM” can work in reality (once we know what it means to individual disciplines)
    - FAIR expert group ++ ?

- Establish policies to ensure that the necessary (scalable, durable, reliable) **infrastructure services** are set up & maintained

- A tail of post-project **funding** – or a home for post-project data (+meta-data+doc+s/w etc.) should be the default

- Support communities in the inevitable service **migrations** (nothing is “eternal”)

CERN

What is?

- Preservation
  - **Data preservation** refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary.
- Curation:
  - **Digital curation** involves maintaining, preserving and **adding value** to digital research data throughout its lifecycle.
- Stewardship:
  - **Even more** – including decisions on what data to preserve, what is the necessary meta-data (and perhaps also data management during active life of the data).
  - (From cradle to grave, according to EU HLEG report claiming a missing 500,000 data scientists)
  - 5% “total project” tax proposed (and disputed by some)

ISO 16363 certification of CERN

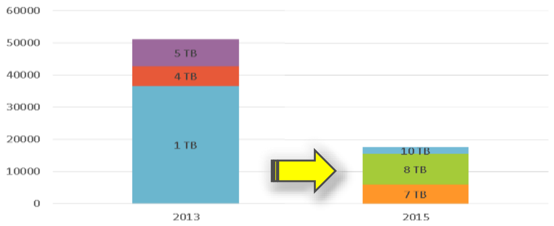
- ISO 16363 follows OAIS breakdown:
  3. **Organisational Infrastructure;**
  4. **Digital Object Management;**
  5. **Infrastructure and Security Risk Management.**

- Many of the elements in 3) and 5) covered by existing (and documented) CERN practices
  - **Some “weak” areas – being addressed – include disaster preparedness / recovery (together with EIROForum)**
  - **And we haven’t really started to address 4) yet…**

➢ **Next step is “stage 1” external audit to high-light those areas requiring attention**
  - **May just be a question of documentation,**
    e.g. CERN is not going to change its financial practices (MTP etc) as a result of ISO 16363!

Bit Preservation: Steps Include

- Controlled media **lifecycle**
  - **Media kept for 2 max. 2 drive generations**



- Regular media **verification**
  - When tape written, filled, every 2 years…
- **Reducing** tape mounts
  - Reduces media wear-out & increases efficiency
- Data **Redundancy**
  - For “smaller” communities, a 2nd copy can be created: separate library in a different building (e.g. LEP – **3 copies at CERN**!)
- **Protecting** the physical link
  - Between disk caches and tape servers
- Protecting the **environment**
  - Dust sensors! (Don’t let users touch tapes)

Constant improvement: reduction in bit-loss rate: 5 x 10^{-16}

See German’s presentation at March DPHEP workshop

Organisational Infrastructure

3.1	Governance & Organisational Viability	Mission Statement, Preservation Policy, Implementation plan(s) etc. Operational Circular, DPHEP Reports
3.2	Organisational Structure & Staffing	Duties, staffing, professional development etc.
3.3	Procedural accountability & preservation policy framework	Designated communities, knowledge bases, policies & reviews, change management, transparency & accountability etc. Generic descriptions refined by project DMPs
3.4	Financial sustainability	Business planning processes, financial practices and procedures etc.
3.5	Contracts, licenses & liabilities	For the digital materials preserved…

Infrastructure & Security Risk Management

5.1	Technical Infrastructure Risk Management	Technology watches, h/w & s/w changes, detection of bit corruption or loss, reporting, security updates, storage media refreshing, change management, critical processes, handling of multiple data copies etc
5.2	Security Risk Management	Security risks (data, systems, personnel, physical plant), disaster preparedness and recovery plans …

Digital Object Management

4.1	Ingest: acquisition of content	
4.2	Ingest: creation of the AIP	Archival Information Package
4.3	Preservation planning	
4.4	AIP Preservation	
4.5	Information management	“FAIR” etc
4.6	Access management	

The plan is to address these after metrics 3 & 5…

Need to agree on scope: only “Open Data”?

Open (Linked) Data

★ Available on the web (whatever format) but with an open license, to be Open Data

★★ Available as machine-readable structured data (e.g. excel instead of image scan of a table)

★★★ as (2) plus non-proprietary format (e.g. CSV instead of excel)

★★★★ All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff

★★★★★ All the above, plus: Link your data to other people’s data to provide context

From https://www.w3.org/DesignIssues/LinkedData.html