ILDG (part 2): Middleware and LDG Aspects

German-Japanese Workshop, Mainz

Hubert Simma
DESY

Sept 27, 2024

Overview

1. FAIR Data
2. Authentication and Authorization
3. Status and Progress of ILDG 2.0
4. Use-cases and Demo
5. Outlook and Challenges

FAIR data: Motivation

Findable
Accessible
Interoperable
Reusable

force11.org
⋮
Wilkinson 2016
⋮
go-fair.org

- ❒ required by funding agencies
- ❒ provide quality standard for data (configs and results)
- ❒ simplify daily work (after some initial effort)
- ❒ allow to give (and receive) credits for shared data
- ❒ help to save resources

EU Commission 2016

➜ guiding principles (not implementation) for scientific data management and stewardship

FAIR data: A local (private) implementation

- Choose some **database** system
  e.g. buy a big disk (with POSIX file system) and use standard tools (`ls`, `grep`, ...)

- Issue **Persistent Identifiers (PID)**:
  e.g. URI-like (not file names!)

- Each data object is an entry (row) in the database (table) with the **fields (columns)**:

∗PID	metadata	data

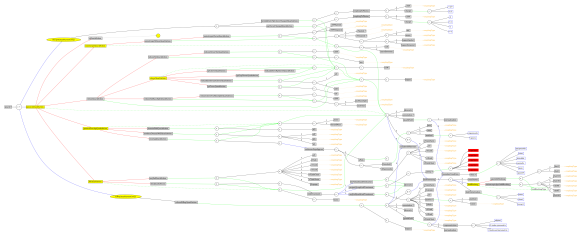
Wilkinson Box 2 ↓

- **Metadata:** should include rich info on → see Hideo's talk (R1)
  - content (F2)
  - provenance (R1.2)
  - related PIDs (F3, I3)
  - ... (e.g. specifications, formats, vocabularies)

FAIR data: Middleware requirements

Metadata must be

- registered or indexed in searchable resource (F4)
  ➜ query language (e.g. SQL, Xpath, JSONPath)
- retrievable by PID (A1)
- accessible even when data is no longer available (A2)
- machine actionable and use a formal language (I1)
  ➜ validation by well-defined and extensible schema (e.g. XML)

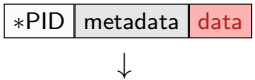


FAIR data: A shared implementation

Requires

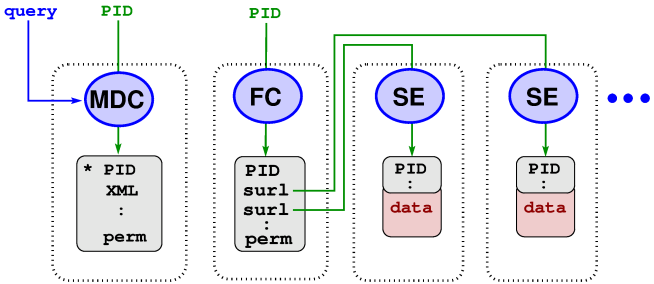
- ❒ additional metadata elements:
  - license (R1.1)
  - permissions (A1.2)
- ❒ minimal **Authentication and Authorization Infrastructure (AAI)**

FAIR data: A distributed implementation



Distinct **web services** (not pages):

- Metadata Catalogue (MDC)
- File Catalogue (FC)
- Storage Element (SE)
- AAI



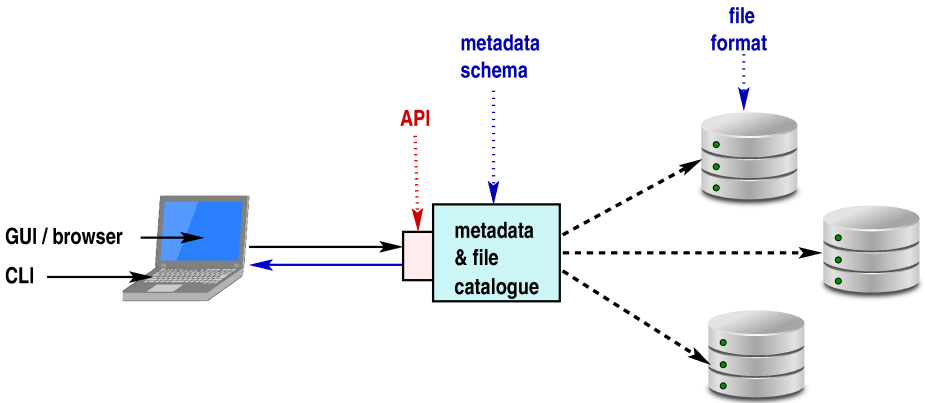
where

- separate MDC and SE becomes mandatory for large data objects (cost of search)
- multiple SE may become mandatory in practice (replication, funding, ownerhip)
- FC becomes mandatory if there are multiple SE or varying storage locations (SURL)

ILDG: 1 MDC and 1 FC in each Regional Grid with standardized API

ILDG: A Federation of autonomous Regional Grids (RG)

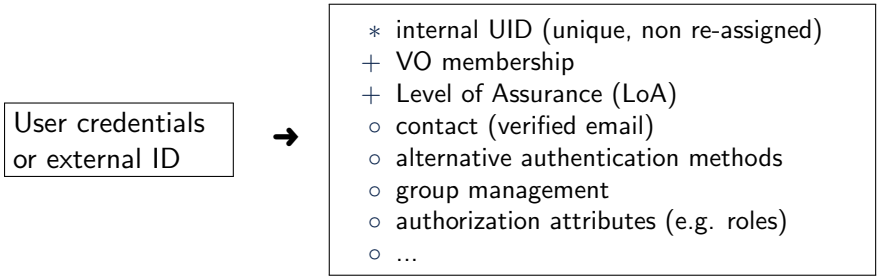
- Interoperable services



- Forming a single registered "Virtual Organization" (VO)

Authentication and Authorization Infrastructure

❒ Provides a single VO-wide user registration (Single Sign On and Attribute Authority)



❒ Typically the AAI acts as token provider (with policy engine)

New AAI: VOMS → IAM

	ILDG 1.0	ILDG 2.0
SSO + Registration	VOMS (DESY)	INDIGO IAM (INFN)
Credential Provider	CA	IdP (or CA)
AuthN transport	X509 proxy certs	ID-token (OpenID Connect / JWT)
Trust Federation / LoA	IGTF	eduGAIN + CoCo or R&S (or IGTF)
AuthZ transport	VOMS proxy cert	Access-token (Oauth2)
AuthZ attributes	only VO membership	fine-grained (at project/resource level)!

N.B.: All VOMS services (WLCG) decommissioned since July 2024.

New Attribute-Based Access Control (ABAC Model)

- **Attributes** of
  - subject (user)
  - action (R/W)
  - object ([meta]data)
  - context
- **Access controlled** by
  - Policy Decision Point → Access Control Service (ACS)
  - Policy Enforcement Points → Resource Servers (RS = MDC, FC, SE)
- Defines a general (possibly huge) **many-to-many relation**

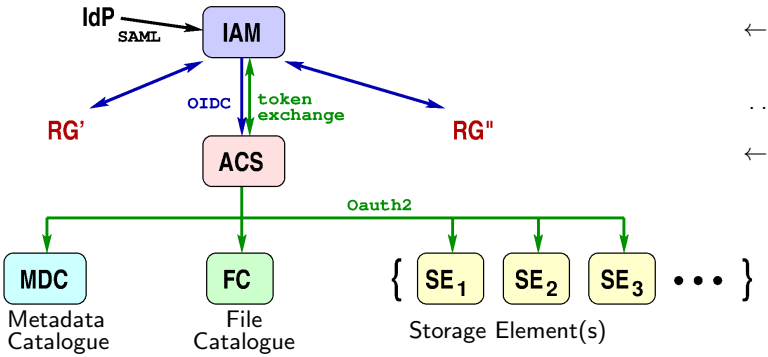
$$\text{user} \longrightarrow\!\!\!\!\!\!\times\!\!\!\!\!\!\longleftarrow \text{object}$$

- Implemented as composition of (smaller) relations

$$\text{user} \xrightarrow[\substack{\text{IAM}\\\text{or ACS}}]{} \text{group} \xrightarrow[\text{ACS}]{} \text{scope} \xrightarrow[\text{RS}]{} \text{object}$$

New ILDG Architecture

- Modular building blocks (reusable in other use-cases, e.g. in PUNCH and beyond)
- Following AARC Blueprint architecture



← Identity and Access Management (AAI)

ILDG

RG

← Access Control Service (Policy Decision Point and token "broker", not issuer)

ILDG solution is in some aspects already ahead of (and possibly taken up by) other experiments (e.g. CMS, ATLAS, Belle, CTA, ALPS, ...)

(Naive) Aim for (I)LDG in 2022

Overhaul of ILDG and RG infrastructure components which were outdated or unusable (except by experts) after >10 years of operation

PUNCH funding for a professional SW developer (Basavaraja BS)

⇓

Re-factor MDC and add new FC with

- up-to-date technology and security (REST, Java)
- simple deployment (Docker containers)
- multiple collections (freely configurable XSD schemata)

⇓

"MDC Reference Implementation for LQCD" (PUNCH deliverable)

Additional Aims and Activities towards ILDG 2.0

- ❒ Foresee support for some user desirables
  - no grid certificates
  - access control
  - markup
  - data publishing
  - tools and docu
  - explore technology trends

➜ **ILDG 2.0**
→ IAM + re-design of MDC and FC
→ new ACS
→ QCDml revision + GUI
→ e.g. 3rd MD schema + RA
→ user support and training
→ object store (cloud)

- ❒ Re-activate and seek support from
  - Board (Frithjof, Yoshinobu)
  - MDWG (Tomoteru, Hideo)
- ❒ Intense contacts with and advice from
  - external experts (PUNCH, Helmholtz, WLCG, IAM and Rucio developers)
  - ILDG pioneers (Tomoteru, Dirk, et al.)

Status of ILDG 2.0

- ✔ No change of basic concepts w.r.t. ILDG 1,
  but drastically changed implementations
- ✔ **IAM** (MoU, AUP, technical+admin+legal issues, final RG setup)
- ✔ **VO Policy** (part of AUP)
- ✔ New **QCDml** revision ready to be released ⇔ enabling new uploads → Hideo
  (CLS, ETMC, HotQCD, JLQCD, openQCD, QCDSF, RCstar, . . .)
- ✔ **Storage Elements**
  - JLDG: 1 (Gfarm, not yet token-enabled, no write access control needed) → Hiroshi
  - LDG: 4 (token + X509 and curl+gfal enabled, Juelich yet "write-only")
- ✘ **MDC, FC, ACS**
  - SW development to be completed in 2024 → Basavaraja
  - Version 1 (certificates) → Version 2 (token-only)
  - several instances deployed (JLDG: 1, LDG: 1+, and UK planned)
  - critical transition (skipping intermediate version with cert + token support)

ILDG Use-cases

- **Consumer of embargoed data**, i.e. only collaboration-internal sharing
  - list/download metadata (e.g. by web page or command line tool, metadata is public)
  - get read permission (from project manager)
  - download configs (requires authentication and authorization)

- **Consumer of "public" data**, i.e. community-wide access (but subject to license)
  - typically need to search relevant data first (powerful Xpath queries)
  - download configs (requires authentication)

ILDG Use-cases (cont'd)

- **Producer of data** (collaboration-internal or community-wide sharing)
  - convert configs (if needed, code-specific tools)
  - pack configs (e.g. by `ildg-binary` tool)
  - markup (e.g. by templates or GUI, easy if part of production workflow!)
  - procure storage space (e.g. assigned to project by RG admin)
  - set public or restricted read permissions (by project manager)
  - get write permission (from project manager)
  - register ensemble metadata (requires authentication and authorization)
  - register and upload configs (" " ")

ILDG Demo

- ❒ Simple listings (web page or wrapper scripts for `curl` → portable "ltools")
- ❒ Xpath queries
- ❒ IAM dashboard
- ❒ Obtain tokens (`oidc-agent` or `curl`)
- ❒ Markup GUI
- ❒ Pack raw config (`ildg-binary`)
- ❒ Upload packed config (`curl` + token)

Outlook (TODO)

- ✻ Completion of new Services (MDC, FC, ACS) ← SW developer
- ✻ Release (and fast minor revisions) of QCDml schemata ← MDWG
- ✻ Specification of new API ← MWWG
- ✘ User tools (clients, CLI, GUI) and documentation (hands-on) ← ALL can contribute!
- ✢ Re-activation of other Regional Grids
- ✢ Support for further binary formats (HDF5)
- ✢ Support for Data Publishing (DOI) and data beyond configs (observables?)
- ✢ OAI-OMH interface for metadata harvesting (e.g. by inspirehep)

Challenges (TBD)

- **ILDG-wide: Sustainability of middleware and services**
  (expertise and person power to maintain services for next 5-10 years)

- **RG-specific: Provisioning of (storage) resources**

Further Discussion Topics (also beyond in this workshop)

- Organization of LDG (EuroLat?)
  - NO board or meetings
  - NO spokes person
  - NO joint efforts for storage resources
  - NO RG-specific policies (resource allocation, embargo periods, publishing, ...)

- Requests/alternatives to ILDG
  - Data publishing (DOIs, Data Repositories)
  - Sharing of data beyond configs (e.g. observables)
  - Definition of an ontology for LQCD

Links

Web pages

- ILDG home (to be improved and moved)
- IAM (user registration)
- MDC interface (simple listings)
- Gitlab (QCDml draft, API, client tools, containerized SW environment, examples, ...)
- Hands-on worhshop
- Markup tool

Email

- `ildg-contact@desy.de`
  (Point of contact to Board or WGs for any questions, requests, suggestions, etc.)
- `ildg-info@desy.de`
  (Moderated mailing list for info on meetings, news, etc.)