This deck of slides is based on the CERIF Tutorial by Brigitte Jörg, CERIF TG Leader 2004–2012, and contains slides by Valérie Brasse, euroCRIS Executive for projects 2011–2015. It also contains slides co-developed by Dragan Ivanović for the CERIF Refactoring Project.
Jan Dvořák

**euroCRIS**
- CERIF TG Leader since 2013
- President since 2023

**Charles University, Institute of Information Studies and Librarianship**
- Researcher & Lecturer

**Czech Technical University in Prague, Computing and Information Centre**
- IS Analyst: V3S+EZOP – operations and further development of the in-house built institutional CRIS

- Leader of the team: Research, Development & Innovation Information System (the national CRIS for [CZ])
- In other roles for the system since 1997
Research Information

The process of research
- Research projects
- Funding
- Research infrastructures

The research actors
- Researchers
- Institutions
- Funders
- Publishers
- Facility operators
- Associations

Relationships

Research results
- Outputs (Publications, Research Datasets, Patents, …)
- Outcomes, Impacts
Who needs Research Information?

- **Researchers**: visibility, finding collaborations, competitors, CV generation
- **Decision Makers**: performance, strategic decisions, priorities, comparisons
- **Funding Organisations**: distribution of programs, evaluation of results, finding reviewers
- **Research Organisations**: integration and interoperability, strategic management
- **Intermediaries / Brokers**: finding research results of potential market or innovative value
- **Enterprises**: finding information for participation in projects, partnerships, usage of results
- **Libraries**: acquisition, dissemination
- **Publishers**: finding reviewers, editors
- **Educators**: integration of relevant findings into lectures and training
- **General Public**: information and education, interest
- **Project Managers**: overview of ongoing activities
- **Media**: distribution and communication
Research Information is heavily interlinked
### Different viewpoints on Research Information

<table>
<thead>
<tr>
<th>Information perspective</th>
<th>Useful for</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researcher, his/her activities, results, ...</td>
<td>Researcher profile / CV system</td>
</tr>
<tr>
<td>Research Project, its consortium, team, funding, results, ...</td>
<td>Project webpage / report</td>
</tr>
<tr>
<td>Organisational Unit, its activities, results, ...</td>
<td>Unit webpage / report / evaluation report</td>
</tr>
<tr>
<td>Publication</td>
<td>Bibliometrics / panel evaluations</td>
</tr>
<tr>
<td>Research Facility</td>
<td>Tracking &amp; reporting usage</td>
</tr>
</tbody>
</table>
What characterizes a research project?

- An acronym
- A name or title
- A code (identifier), for example a Grant number
- A short or long description (abstract)
- A (planned) start date
- A (planned) end date or duration
- [A source of funding]
- [A few scientific publications]
- [A project coordinator]
- [A research domain]

Source: http://cordis.europa.eu/project/rcn/106635_en.html
A name or title

A short or long description (abstract)

A code (identifier), for ex a Grant number

A web page (URI)

A (planned) start date
A (planned) end date or duration

[A project coordinator]

[A source of funding]

[A research domain]

A few keywords

Source: http://gtr.rcuk.ac.uk/project/A49CA721-687A-4D55-8FDF-9B60375B6EA8
The PROJECT entity has properties (attributes) and is linked to other entities.

The multilingual attributes are represented by a linked entity each.

* “start date” and “end date” are deprecated in v1.6
CERIF naming rule: in English, abbreviated, starting with *cf*

*Example:* Project title = *cfProjTitle*
Example in DB

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Table</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>cfProjId</td>
<td>cfProj</td>
<td>base</td>
</tr>
<tr>
<td>cfAcro</td>
<td>cfProj</td>
<td>base</td>
</tr>
<tr>
<td>cfURI</td>
<td>cfProj</td>
<td>base</td>
</tr>
<tr>
<td>cfStartDate</td>
<td>cfProj</td>
<td>base</td>
</tr>
<tr>
<td>cfEndDate</td>
<td>cfProj</td>
<td>base</td>
</tr>
<tr>
<td>cfLangCode</td>
<td>cfProj</td>
<td>base</td>
</tr>
<tr>
<td>cfTrans</td>
<td>cfProj</td>
<td>base</td>
</tr>
<tr>
<td>cfProjTitle</td>
<td>cfProjTitle</td>
<td>lang[en,o]</td>
</tr>
<tr>
<td>cfProjKeyw</td>
<td>cfProjKeyw</td>
<td>lang[de,h]</td>
</tr>
<tr>
<td>cfProjAbstr</td>
<td>cfProjAbstr</td>
<td>lang</td>
</tr>
</tbody>
</table>

Project definition

A temporary endeavor undertaken to create a unique product, service or result.
Source: the Project Management Institute, https://www.pmi.org/about/learn-about-pmi/what-is-project-management

In the research information domain, one typically tracks:
(1) research projects, where the result is an addition to the body of knowledge of mankind,
(2) technology development projects, where the result is a particular technology or product,
(3) innovation projects, where the result is an improvement of a product or process, and
(4) projects that create or enhance infrastructure for research, technology development or innovation.

Depending on the scope one can also track finer levels of granularity: stages, work packages, sometimes even down to individual tasks. All such activities are also modelled using the Project entity and linked using the recursive link relationship.

The Project entity only captures details of the project scope and plan. Information about the resources needed to execute the project such as the funding (i.e., the grants received), the people and organisations involved, the supporting infrastructures, the outputs produced, etc. is contained in separate entities (the Funding entity, the Person entity, the OrgUnit entity, the infrastructure entities, the result entities respectively) and is linked to the Project.
INTERMEDIATE SUMMARY

• CERIF is:
  – A conceptual model
  – A storage format in relational database
  – A set of exchange formats (XML, Linked Data)

• CERIF supports multilingual data, storing the original value of a literal attribute, and for any other language, a value translated by a machine and/or a human

• So far, we have seen the CERIF Entity “PROJECT” (cfProj)
We have seen how to represent, store or exchange metadata about research projects.

Similarly:

- What characterises a person (researcher, Ph.D.,...)?

- What characterises an organisation (research laboratory, institute,...)?
CERIF Base Entities

- Person
- Project
- OrganisationUnit
Person definition

A human being as an individual.
Source: https://en.oxforddictionaries.com/definition/person

The kind of involvement of a Person in the research ecosystem is specified in the links with the organisations, the services, etc. This typically includes:
(1) **researchers** (Persons performing research in an Organisation Unit as employees or students);
(2) **authors** and **contributors** (Persons signing a publication, creators of data sets, software developers, etc.);
(3) **investigators** and project **participants** (Persons involved in a Project as principal investigators, co investigators, project managers, consultants, etc.);
(4) **management** (directors, rectors, deans, department heads, etc.);
(5) **support** staffs (technicians, responsible for Equipment, librarians and digital asset curators, administrative staff, etc.).
One Person typically has many of these relationships.
OrgUnit definition

Organisation Unit: an organisation, a unit therein, a committee or any other group of people that has a collective goal. Organisation Units are not necessarily formalized as legal entities.

In the research information domain Organisation Units typically represents:
(1) organisations that perform research (universities, research institutes, corporations) and their subdivisions (faculties, schools, departments, research groups) and other associated bodies (boards, advisory bodies);
(2) organisations that fund research (funders, their divisions and evaluation panels);
(3) scientific associations and networks;
(4) publishers, facility operators and other service providers in the research space;
(5) authorities, such as patent offices and standardization or supervision bodies; and
(6) other bodies: editorial boards, evaluation panels, or committees of all kinds.
Common European Research Information Format
CERIF Result Entities

- ResultPublication
- ResultPatent
- ResultProduct
CERIF Infrastructure Entities

- Facility
- Equipment
- Service
CERIF Infrastructure Entities

**Facility**
- ID
- Acronym
- URI
- Title
- Description
- Keywords
- Type
- Status
- Subject

**Equipment**
- ID
- Acronym
- URI
- Title
- Description
- Keywords
- Type
- Status
- Subject

**Service**
- ID
- Acronym
- URI
- Title
- Description
- Keywords
- Type
- Status
- Subject
CERIF General Pattern

A typical CERIF entity has:
• Identifier (internal)
• Attributes
  • the basic ones
  • the multi-lingual ones
• External Identifiers
• Classifications
  • Type
  • Status
  • Subject area
  • + other
• Links
  • to other entities
  • recursive, a.k.a. self-referential
CERIF 1.6
Some CERIF Link Entities

- Person_ResultPublication
  - role=author

- OrganisationUnit_ResultPublication
  - role=author's affiliation

- Project_ResultPublication
  - role=deliverable

- Person_OrganisationUnit
  - role=research assistant

- Project_OrganisationUnit
  - role=coordinator

- Person_Project
  - role=principal investigator

- Project

- ResultPublication
Another example
(slides by Keith Jeffery)
Generic Linking Entity Structure

Base object 1
(FK)

role: cfClassification
(FK)

cfClassId
cfClassSchemeId

cfStartDate
cfEndDate

Fraction (optional)

Proportion or Intensity

Base object 2
(FK)
Recording Change in CERIF

Example: The Principal Investigator of project $P$ changes: $X$ is replaced by $Y$ effective date $D$.

Before:
- **Validity range**: $-\infty \ldots +\infty$
- **Role**: Principal Investigator : cfClassification
- **Principal Investigator**: $X$

After:
- **Validity range**: $-\infty \ldots D$
- **Role**: Principal Investigator : cfClassification
- **Principal Investigator**: $X$
- **Validity range**: $D \ldots +\infty$
- **Role**: Principal Investigator : cfClassification
- **Principal Investigator**: $Y$
CERIF Federated Identifiers

• Records the “tag” by which an object is known elsewhere
• For any CERIF research entity
• “Identifier Types” classification scheme
• (optionally) Connects to a Service representing the issuer of the identifier
  • Usually an information system
CERIF Federated Identifiers: Examples

**Publication**
- DOI
- PubMed Central ID
- WoS UT, Scopus EID
- ISBN, ISSN

**Person**
- ORCID iD
- ISNI
- ResearcherID, Scopus Author ID
- IdRef, DAI, Número Lattes
- Social Security Number or another administrative identifier
- Staff Id in an HR system

**Project/Grant**
- Funder’s reference number
- Organisation’s reference number

**Organisation**
- ROR ID
- FundRefID
- Wikidata reference
- ISNI
- RingGold number
- Org Enhanced (WoS)
- VAT Identification Number or another administrative identifier

**Classifications**: External Codes
CERIF Semantic Layer

Central place to store declared semantic classifications

Allows to capture any Schema or Structure
- Flat Lists
- Thesauri
- Classification Systems (e.g. SKOS, ...)
- Taxonomies
- Ontologies

Open / Extensible in all directions
- New Schemas
- New Concepts / Terms
- New Relationships

Enables to manage
- Roles, Types, Statuses, ... Semantics
- Subject Headings
- Versioning and archiving (start+end dates)

Allows for Mappings between Schemes
- skos:closeMatch, skos:exactMatch or any other mapping you need
Metadata Layers

Discovery metadata
DC, VIVO, MODS, METS, eGMS, DCAT, ...

Contextual metadata
CERIF

Detailed metadata
Domain-specific standards
Current Research Information System

- Ongoing
- Past, of current interest
- Currently planned or decided upon
CERIF development

By the CERIF Task Group of euroCRIS

Adopting open-source software projects tools & best practices:

→ https://github.com/EuroCRIS/CERIF-DataModel
→ https://github.com/EuroCRIS/CERIF-Vocabularies
→ CC BY license

Two branches:
- master: latest official release (1.6.1)
- develop: on-going development
### Basic Information

<table>
<thead>
<tr>
<th>Project</th>
<th>CERIF: the Common European Research Information Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>CERIF Data Model</td>
</tr>
<tr>
<td>Version</td>
<td>CERIF current development DM</td>
</tr>
<tr>
<td>Company</td>
<td>euroCRIS</td>
</tr>
<tr>
<td>Author</td>
<td>CERIF Task Group</td>
</tr>
<tr>
<td>Created</td>
<td>30.6.2014</td>
</tr>
<tr>
<td>Last Modified</td>
<td>22.12.2017</td>
</tr>
</tbody>
</table>

### Statistic Information

<table>
<thead>
<tr>
<th>Entities</th>
<th>295</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attributes</td>
<td>1829</td>
</tr>
<tr>
<td>Domains</td>
<td>1</td>
</tr>
<tr>
<td>User Data Types</td>
<td>0</td>
</tr>
<tr>
<td>Relationships</td>
<td>671</td>
</tr>
<tr>
<td>Indexes</td>
<td>0</td>
</tr>
<tr>
<td>Keys</td>
<td>298</td>
</tr>
<tr>
<td>Views</td>
<td>0</td>
</tr>
<tr>
<td>View Relationships</td>
<td>0</td>
</tr>
<tr>
<td>Procedures</td>
<td>0</td>
</tr>
<tr>
<td>Defaults</td>
<td>0</td>
</tr>
<tr>
<td>Schemas</td>
<td>0</td>
</tr>
<tr>
<td>Users</td>
<td>0</td>
</tr>
<tr>
<td>User Groups</td>
<td>0</td>
</tr>
<tr>
<td>Notes</td>
<td>0</td>
</tr>
</tbody>
</table>
Common European Research Information Format

- CERIF is an EU Recommendation to Member States
- The European Commission (EC) has authorised euroCRIS to maintain and develop CERIF and its usage

http://cordis.europa.eu/cerif/
INTEROPERABILITY
Interoperability

1. Agreed Semantics
2. Agreed Format
3. Agreed Protocol
CERIF-XML exchange formats (based on XML Schema)

**Original**
1:1 with the ER structure
Only uses embedding for multilingual texts
→ Many foreign key relationships
→ Takes several API requests to get a presentable form of an object

**2nd Generation**
Template XML Schema → adaptation
Profiles: Useful subsets of CERIF for specific research information exchange scenarios

1. Specify a subset of CERIF entities & attributes
2. Fix semantic vocabularies to use
3. Add integrity constraints

Ex.: OpenAIRE Guidelines for CRIS Managers 1.0.

Ex.: OpenAIRE Guidelines for CRIS Managers 1.1.
OpenAIRE Guidelines for CRIS Managers 1.1


- Introduction

- CRIS Information elements relevant for OpenAIRE
  - CERIF, CERIF XML

- Technical implementation guidelines
  - CERIF XML as payload of OAI-PMH 2.0

- Version 1.1.2 being finalized
  https://github.com/openaire/guidelines-cris-managers
  with Andreas Czerniak and Dragan Ivanović
Version 1.1: Aligned vocabularies

- COAR Resource Types (v. 3.1)
- COAR Access Rights
- ISSN Media List
Information scope of
the OpenAIRE Guidelines for CRIS Managers 1.1

The Core:
- The CRIS as a Service
- Organisations
- Researchers

Activities:
- Projects
- Funding
- Events

Outputs:
- Publications
- Datasets, Software & other Products
- Patents

Infrastructure:
- Equipment
Identifiers

- **Publication:**
  - DOI
  - Handle
  - PMCID
  - ISI Number (WoS Accession Number)
  - SCP Number
  - ISSN
  - ISBN
  - URL
  - URN

- **Persons:**
  - ORCID iD
  - ISNI
  - ResearcherID
  - Scopus Author ID

- **Products:**
  - DOI
  - ARK
  - Handle
  - URL
  - URN

Plus a generic option
Supporting artifacts

• XML Schema
  – Namespace https://www.openaire.eu/cerif-profile/1.1/
  – A few Schematron integrity rules embedded

• A comprehensive set of examples

• Prototype validator
A Research Graph:
The example set from the OpenAIRE Guidelines for CRIS Managers 1.1
Strong selection against hybrids maintains a narrow contact zone between morphologically cryptic lineages in a rainforest lizard
<Publication xmlns="https://www.openaire.eu/cerif-profile/1.1/" id="852734">

  <!-- [ ... ] -->
  <Authors>
    <!-- [ ... ] -->
    <Author>
      <Person>
        <PersonName>
          <FamilyNames>Singhal</FamilyNames>
          <FirstNames>Sonal</FirstNames>
        </PersonName>
      </Person>
      <Affiliation>
        <OrgUnit id="301248">
          <Name xml:lang="en">Museum of Vertebrate Zoology</Name>
          <PartOf>
            <OrgUnit id="329384">
              <Name xml:lang="en">University of California, Berkeley</Name>
            </OrgUnit>
          </PartOf>
        </OrgUnit>
      </Affiliation>
    </Author>
    <!-- [ ... ] -->
  </Authors>
  <!-- [ ... ] -->
</Publication>
Phenotypically cryptic lineages comprise an important yet understudied part of biodiversity;
CERIF Profiles

Producers know what to include

Consumers know what to expect
CERIF Profiles

Useful subsets of CERIF for specific research information interchange scenarios

- Entities & attributes:
  - Profile $\subset$ CERIF

- Semantic vocabularies:
  - Profile – specific choices
  - Sources: CERIF & beyond

- Integrity constraints:
  - Profile $\supset$ CERIF

Profile data is CERIF
CERIF Refactoring

• Address shortcomings of current CERIF

• Modernize CERIF
  – Take away the perceived complexity of CERIF
  – Involve the community in further development of CERIF
  – New modelling approach (and tool), modularity, modern serialization formats

• Ongoing project started 2021
CERIF REFACTURING
CERIF Refactoring Goals

1. Keep the strong features of current CERIF
   a. Multilinguality

2. Take away the perceived complexity of CERIF
   a. Emphasize the conceptual model
   b. Improve documentation

3. Modernize CERIF
   a. Change the modeling notation
   b. Adapt CERIF for usage in APIs and for Linked Open Data
   c. Modern serialization formats
   d. Allow for systematic provenance tracking and verifiable credentials

4. Involve the community in further development of CERIF
   a. Modularity
   b. Open source sw development practices
CERIF Refactoring Means

1. New modelling approach (and tool)
   a. Improved documentation
   b. Element addressability

2. Modularity (⇒ scalability of further CERIF development)
   a. The Core and pluggable modules
   b. Datatypes to represent reusable functional groups
   c. Developed in a decentralized way

3. Modern serialization formats (with appropriate schema expressions)
   a. JSON-LD (JSON Schema, SHACL) for APIs
   b. RDF (RDF Schema, SHACL) for LOD
   c. previously: XML (XML Schema, Schematron)
CERIF Core

Contains:

- common data types
- abstract entities
- commonly used entities
- common constructs

The basis for:

- Particular research information interchange scenarios
- Extensions in terms of coverage
A prototype CERIF Core repository at [https://github.com/EuroCRIS/CERIF-Core](https://github.com/EuroCRIS/CERIF-Core)

CERIF Core

CERIF (=Common European Research Information Format) covers the domain of research information with a focus on the administrative and organizational aspects. It aims to provide machine-processable.
CERIF **Agent**

Def: Anything that has the ability to perform actions. Generalizes the former cfPerson + cfOrgUnit

Subclassed as:

- **Person**
- **Group_or_Organisation_Unit**
  - Organisation_Unit (formalized)
  - Group (has members)

(following **FOAF**)
The Agent Hierarchy
CERIF **Activity**

Def: Anything an Agent does.

Subclassed as:

- **Involvement** *(Membership, Employment or Education)* in a **Group** or **Organisation Unit**
- **Contribution** to a **Document**, a **Project**, an **Event**, …

This is a refinement of the Linking Entities of current CERIF starting from cfPerson or cfOrgUnit *(dateRange is an attribute of Activity)*
Activity hierarchy
CERIF Affiliation Statement

To give context to the Agent in his/her/its Activities.

Person is member OrgUnit

Person with affiliation to OrgUnit is member

OrgUnit is member, represented by Person

OrgUnit is member, represented by Position

OrgUnit is member

Where does the person come from? Why are they there?

Who represents the orgunit? Who should we talk to?
CERIF **Project** & **Funding**

Project: A temporary endeavor undertaken to create a unique product, service or result.  
(Source: the Project Management Institute)

Funding: Money provided by an organization or government for a particular purpose. 
(Source: the Oxford Learner’s Dictionary of Academic English)

Common confusion: funders, institutions and researchers alike may say “Project” when they mean “Project with Funding” or “Funding”.
Funding: a specific type of **Resource**

It's not just Funding that enables research. There is also:

- The expertise, time & effort of Researchers
- Prior knowledge (from Documents)
- Housing the research, the administrative backing by Institutions
- Services offered by Facilities:
  - Laboratories
  - Supercomputing centres
  - Data archives
  - Libraries
  - ...
- **Other Services**
  - Publishing

Most of these can be had in exchange for some Funding
Resources exist in different modes

Planned

Constructed / Secured

Offered / Invitation to use

Considered

Applied for / Negotiated

Approved / Agreed upon

Contracted

Used

Sought for
CERIF **Document**

Def: A document is a bounded representation of a body of information designed with the capacity (and usually intent) to communicate. A document may manifest symbolic, diagrammatic or sensory-representational information.


Accommodates current cfResultPublication, cfResultPatent and some of cfResultProduct
CERIF Core “View from the Top”

Agent -> Activity

Activity -> Document
Activity -> Project
Activity -> Event

Resource
CERIF Modules

Add functionality

- E.g. the Scholarly publication module

Customize the model for a specific scenario

- E.g. institutions to report publications produced with Horizon Europe funding to the EC (tentative)

Contributed by another group of people

Reuse & Extend stuff from the Core and other modules

Candidate modules:

- Dataset metadata
- Research software metadata
- Bibliometric indicators
- Project reporting
- Research Infrastructure description
CERIF Scholarly Publication Module

1st module started along with the Core

Scope:

- Journal Article
- Journal
- Monograph
- (Book Chapter)
- (Proceedings Paper)
CERIF Scholarly Publications Module

A module to represent scholarly publications in the refactored CERIF. Its secondary purpose is to demonstrate the liaison between the CERIF Core and a module.

Status
Thank you!