University of St Andrews
Scotland's first university

600 YEARS
1413 – 2013

CERIF 4 DATASETS

CERIF for Datasets:

Linking and contextualising publications and datasets, and much more …

Scott Brander, Anna Clements, Valerie McCutcheon, Paul Cranner, Ryan Henderson, Kevin Ginty

First Workshop on "Linking and Contextualizing Publications and Datasets", 26th Sep 2013
17th International Conference on Theory and Practice of Digital Libraries : TPDL2013
Sep 22-26, 2013, Valletta,Malta

University of St Andrews
Scotland's first university

600 YEARS
1413 – 2013

CERIF 4 DATASETS

Many thanks to Nikos Houssos for presenting on our behalf

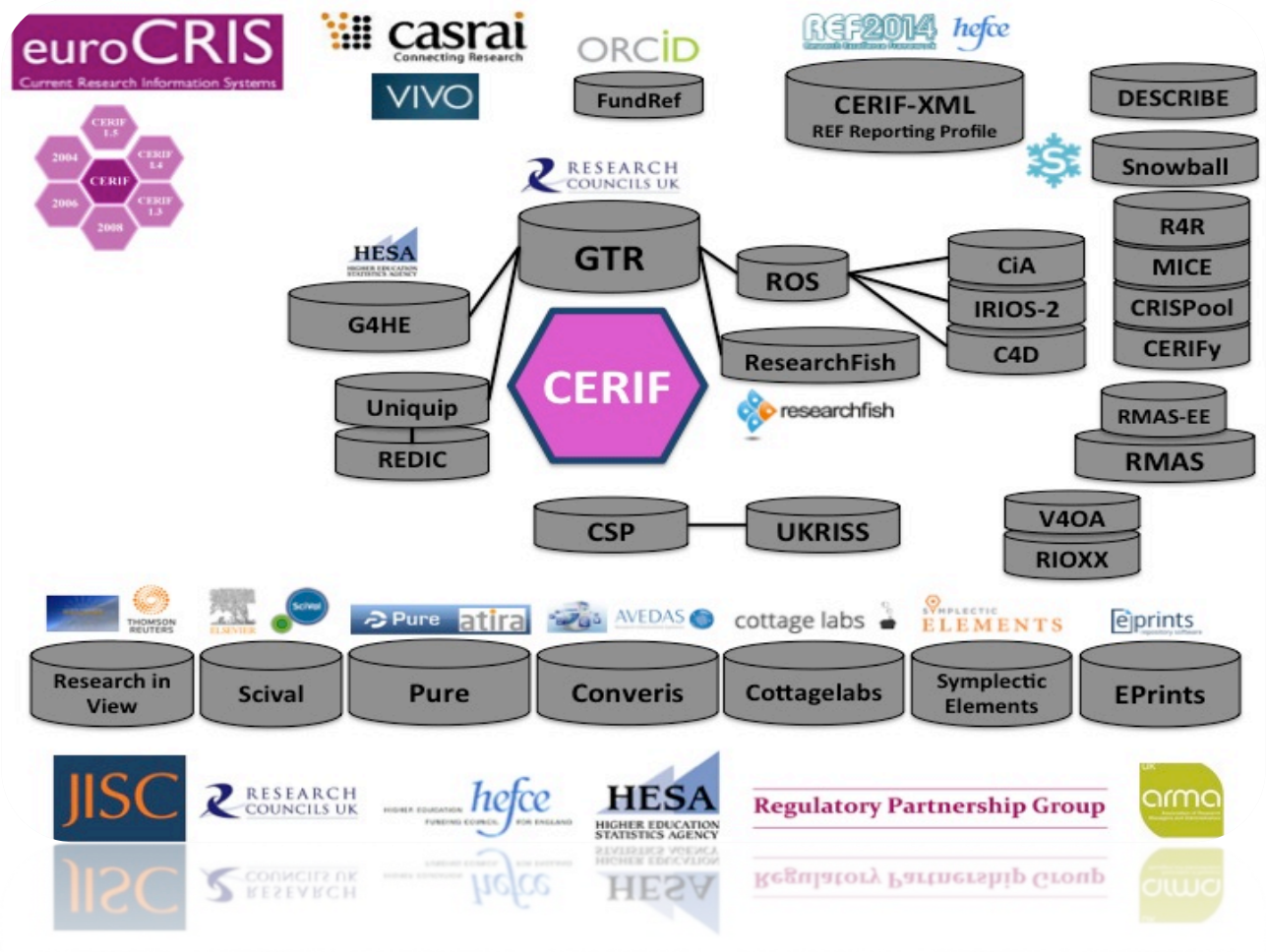
Slides prepared by Anna Clements with thanks to euroCRIS colleagues:
Keith Jeffery, Brigitte Joerg and Jan Dvorak

C4D Summary

- JISC Managing Research Data Programme
- Consortium : Sunderland, Glasgow, St Andrews, NERC, EPSRC, DCC and euroCRIS
- “CERIFication” of the *metadata* about research datasets
- Focus on MEDIN* standard : NERC requirement for http://www.bodc.ac.uk/

* http://www.oceannet.org/

In the UK : The CERIF landscape



CERIF basics

- **C**ommon **E**uropean **R**esearch **I**nformation **F**ormat
- A conceptual model for describing the complete research domain
- A standard for the development, implementation and interoperability of current research information systems (CRIS) and their various application
- Est. 1991; maintained by www.euroCRIS.org
- Ongoing work with OpenAire, DataCite, RD-Alliance, ORCID

euroCRIS basics

- Not for profit organisation of experts
  - Research organisations; funders; publishers; systems providers; standards organisations
- 109 institutional, 38 personal & 20 affiliate members (*euroCRIS annual report 2012)*
- 41 countries; not just Europe
- Main activity : development, maintenance and implementation of CERIF
- Multiple strategic partners, e.g. VIVO, COAR, CODATA, CASRAI, and others

CERIF evolution

CERIF 91

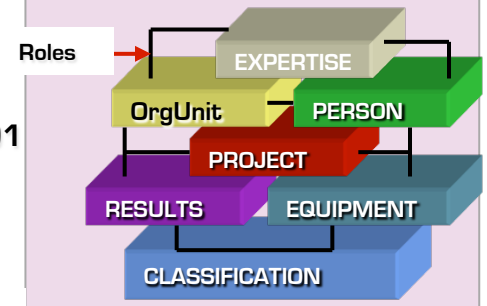
Acronym : ERGO
Participants : Keith Jefffery, Anne Asserson, Rutherford Appleton Lab, Univ Bergen,, many more

- Networking of DBs
- Exchange of Records

- EC Recommendation to Member States

1991

CERIF 2000 Model

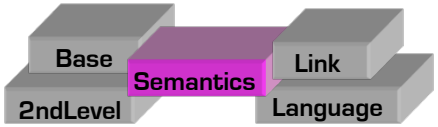


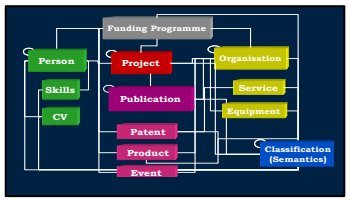
- Data Model
- Multilinguality
- Controlled Vocabulary
- Roles / Types
- User-driven

- EC Recommendation to Member States

2000

CERIF 2006 / 2008 Model





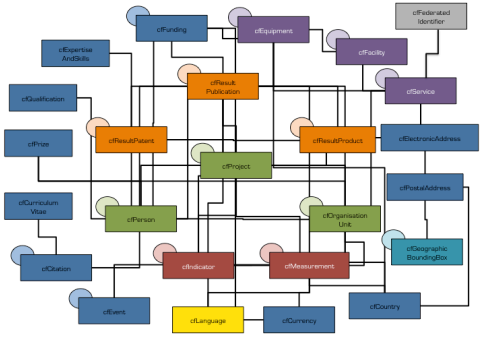
- Data Model
- Model Normalization
  - Robust/Consistent Structure
  - Extensible Structure
  - Semantic Layer
- XML Exchange Specification
- Elaboration on Publication
- CERIF Core Semantics (2008 1.2)

euroCRIS
Current Research Information Systems

2002

2006

CERIF 1.5
CERIF 1.4 (XML)
CERIF 1.3



–Data Model
– Infrastructure
  - Facility, Equipment, Service
- Measurement & Indicator
  - Entities and Link Tables
- Geographic Bounding Box
- CERIF 1.3 Vocabulary
  - UUIDs
  - Terms
  - Schemes
- CERIF 1.4 new XML format
- CERIF 1.5 Federated Identifiers

+ Linked Data

2012

CERIF 1.6

–Data Model
– C4D datasets

2013

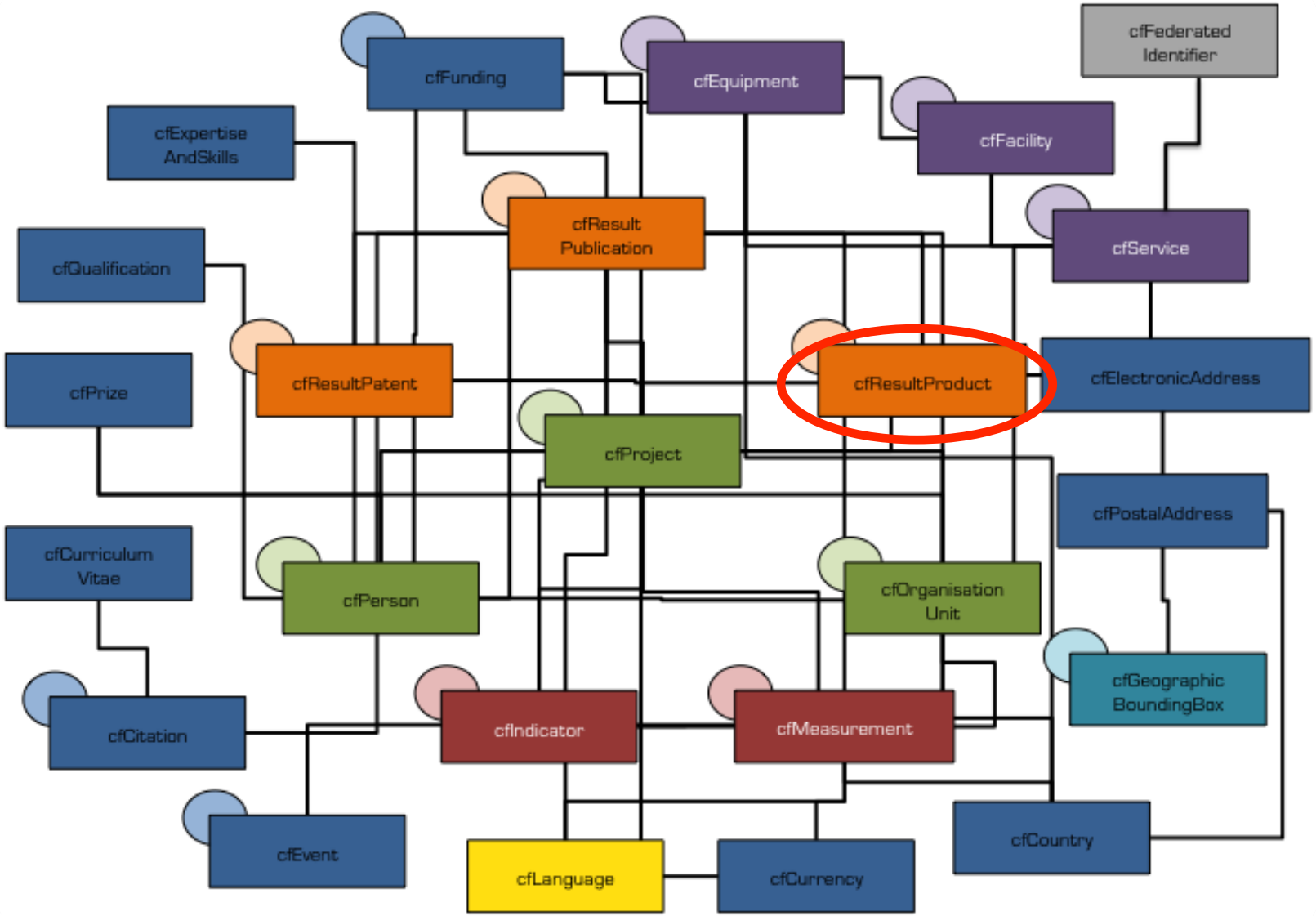
CERIF components

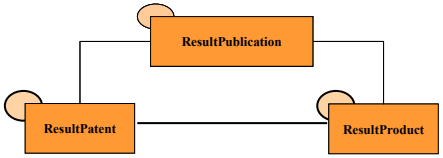
CERIF Entity Types

- Base Entities
- Result Entities
- Infrastructure Entities
- 2nd Level Entities
- Link Entities

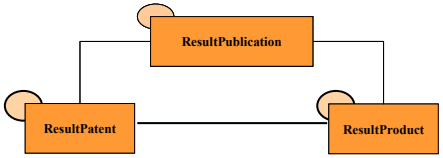
CERIF Features

- Multiple Language
- Semantics
- Measures & Indicators
- Geographic Bounding Box


cfFunding
cfEquipment
cfFederated Identifier
cfExpertise AndSkills
cfFacility
cfResult Publication
cfService
cfQualification
cfResultPatent
cfResultProduct
cfElectronicAddress
cfPrize
cfProject
cfPostalAddress
cfCurriculum Vitae
cfPerson
cfOrganisation Unit
cfGeographic BoundingBox
cfCitation
cfIndicator
cfMeasurement
cfEvent
cfLanguage
cfCurrency
cfCountry


ResultPublication
ResultPatent
ResultProduct

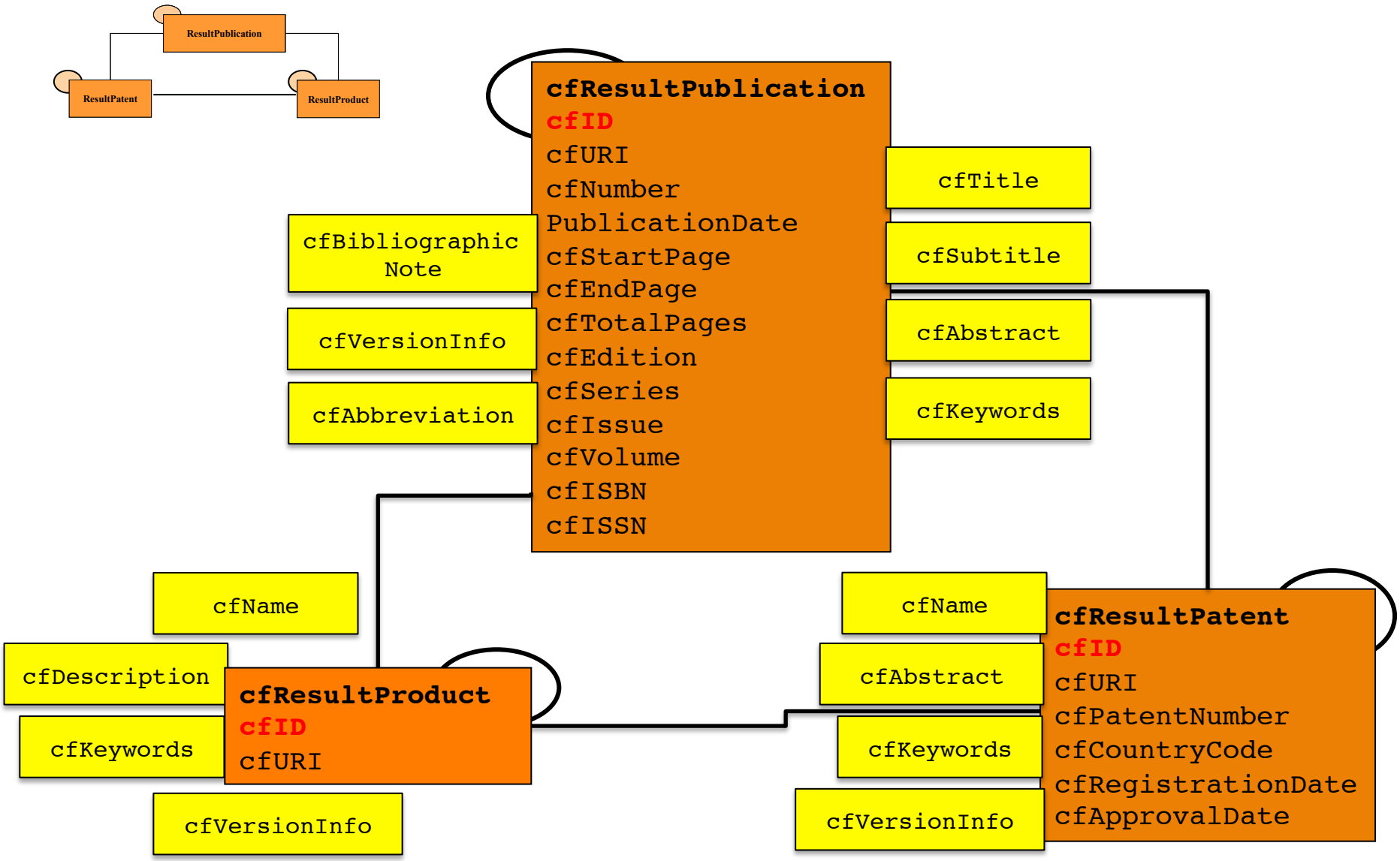

ResultPublication
ResultPatent
ResultProduct



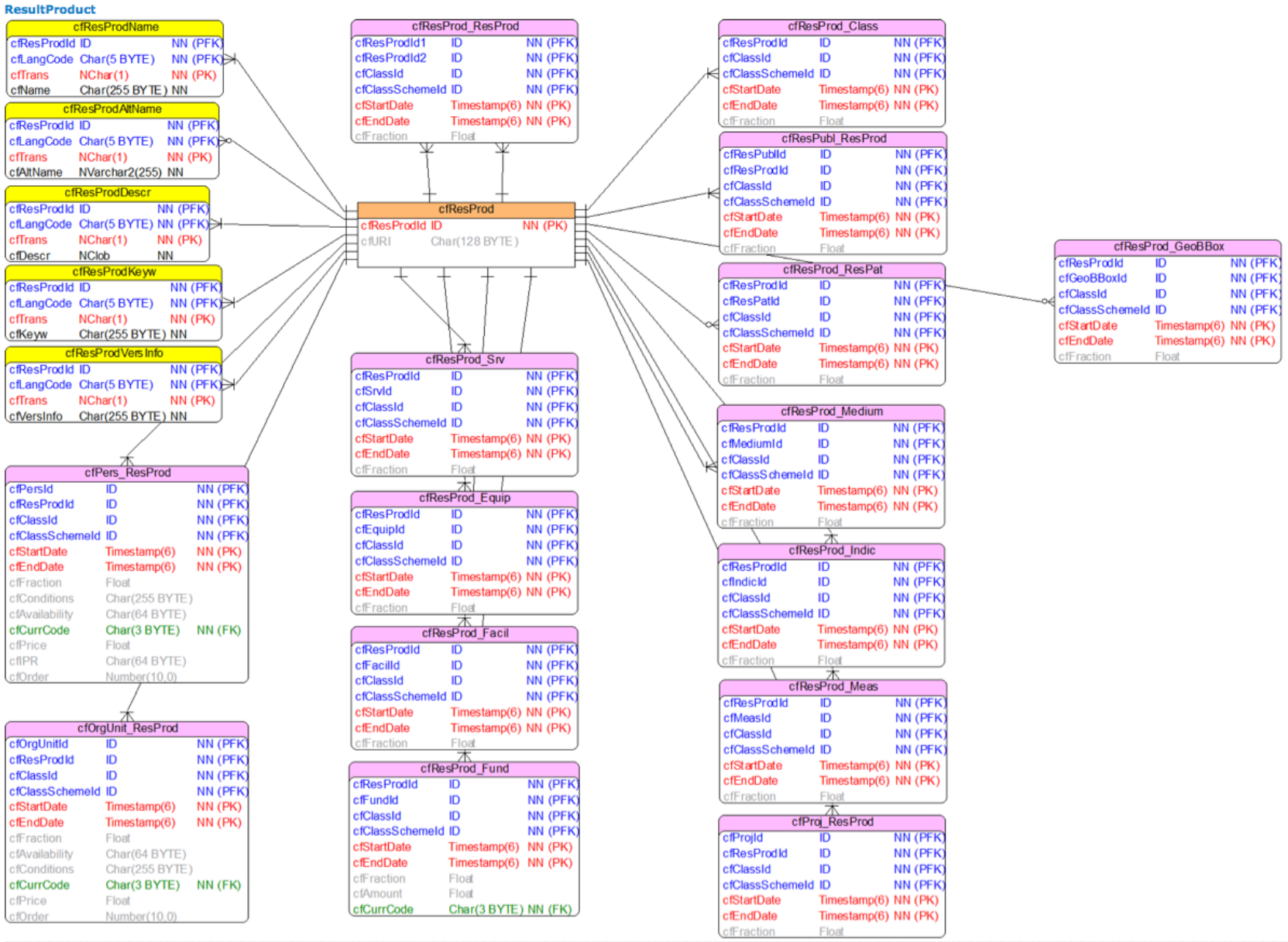
ResultPublication
ID
URI
Title
Subtitle
Abstract
Bibl. Note
PublicationDate
TotalPages
StartPage
EndPage
Keywords

ResultPatent
ID
URI
PatentNumber
Title
CountryCode
RegistrationDate
ApprovalDate
Description
Keywords

ResultProduct
ID
URI


ResultPublication
ResultPatent
ResultProduct
cfResultPublication
cfID
cfURI
cfNumber
PublicationDate
cfStartPage
cfEndPage
cfTotalPages
cfEdition
cfSeries
cfIssue
cfVolume
cfISBN
cfISSN
cfTitle
cfSubtitle
cfAbstract
cfKeywords
cfBibliographic Note
cfVersionInfo
cfAbbreviation
cfName
cfDescription
cfKeywords
cfResultProduct
cfID
cfURI
cfVersionInfo
cfName
cfAbstract
cfKeywords
cfVersionInfo
cfResultPatent
cfID
cfURI
cfPatentNumber
cfCountryCode
cfRegistrationDate
cfApprovalDate

CERIF 1.6 ER cfResProd



Mapping MEDIN to CERIF 1.5

	MEDIN	DataCite v3.0 Mandatory Recommended Optional	CERIF v1.5	Notes
0	Identifier	M	cfResProdId	
1	Resource Title	M	cfResProd, cfResProdName.cfName	
2	Alternative Resource Title		Not supported – proposed to CERIF Task Group	Approved and due v1.6, summer 2013
3	Resource Abstract	R (Description)	cfResProd.cfResProdDescr.cfDescr	
4	Resource Type	R	cfResProd.cfResProd_Class	
5	Resource Locator		cfResProd_Srv.Srvid	
6	Unique Resource Identifier		cfResProd.URI	
7	Coupled Resource		cfResProd_ResProd.classId	
8	Resource Language	O	cfResProd_Class.cfLang with appropriate cfLangCodes	
9	Topic Category	R (Subject)	cfResProd_Class.cfClassId with appropriate classification scheme	
10	Spatial Data Service Type		cfResPubl_Srv.cfClassId linked to cfResProd	
11	Keywords		cfResProd.ResProdKeyw.Keyw and cfResProd.cfResProd_Class.cfClassId	
12	Geographic Bounding Box	R (GeoLocation)	cfResProd_GeoBBox.GeoBoxId	
13	Extent		cfResProd_Class.cfClassId	
14	Vertical Extent Information		Not supported in CERIF. CERIF has a GeoBBox element which can be used to record these attribute, but there is currently no cfResProd_GeoBBox linking element.	Approved and due in v1.6, summer 2013
15	Spatial Reference		cfResProd_Class.cfClassSchemeId	
16	Temporal Reference	M (Publication Year) R (other dates e.g. period of collection)	cfResProd_Class.cfClassSchemeId with temporal reference classification scheme	
17	Lineage		Not currently supported – proposed	CERIF TG still discussing
18	Spatial Resolution		Not currently supported – proposed	Recommendation is cfResProd_GeoBBox
19	Additional Information		cfResProd_cfResPubl.ResPublId with classification scheme	
20	Limitations on Public Access	O (Rights)	cfResProd_Class with appropriate classification scheme	
21	Conditions applying for access and use	O (Rights)	Not currently supported – proposed free text	CERIF TG still discussing
22	Responsible party	M (Creator, Publisher) O (Contributor)	cfOrgUnit_ResProd.OrgUnitId cfPers_ResProd.PersId	
23	Data Format	O	cfResProd.cfResProd_Class with Data Format classification scheme	
24	Frequency of Update		cfResProd_Class.ClassId with Frequency of Update classification scheme	
25	Conformity		cfResProd_Measurement.MeasId	
26	Metadata Date		This is managed by the application	
27	Metadata Standard Name		No recommendation by CERIF Task Group, so was mapped to cfOrgUnit_cfresProd with linking roles	
28	Metadata Standard Version		As per 27	
29	Metadata Language		Is cfLang entity but no link to cfResProd currently	CERIF TG still discussing
30	Parent ID	R (Related Identifier)	cfResProd_ResProd with appropriate classification scheme	

Representing temporal information : TG discussion

Element No.	Element Name	CERIF Entity	Vocabularies/Comments
16	Temporal Reference	cfResProd.cfResProd_Class.cfClassSchemeId	Temporal Extent Scheme (The period the data is related to)

C4D CERIF

```
<cfResProd>
        <cfResProdId>02086348-7c61-4be1-8976-2003afc66052</cfResProdId>
        <cfResProd_Class>
                <cfClassId>temporal_extent</cfClassId>
                <cfClassSchemeId>class_scheme_resultProduct_classification_temporalReference</cfClassSchemeId>
                <cfStartDate>2010-01-01T00:00:00</cfStartDate>
                <cfEndDate>2010-01-02T00:00:00</cfEndDate>
        </cfResProd_Class>
        <cfResProd_Class>
                <cfClassId>publication</cfClassId>
                <cfClassSchemeId>class_scheme_resultProduct_classification_temporalReference</cfClassSchemeId>
                <cfStartDate>2010-01-03T00:00:00</cfStartDate>
        </cfResProd_Class>
        <cfResProd_Class>
                <cfClassId>revision</cfClassId>
                <cfClassSchemeId>class_scheme_resultProduct_classification_temporalReference</cfClassSchemeId>
                <cfStartDate>2010-01-04T00:00:00</cfStartDate>
        </cfResProd_Class>
        <cfResProd_Class>
                <cfClassId>creation</cfClassId>
                <cfClassSchemeId>class_scheme_resultProduct_classification_temporalReference</cfClassSchemeId>
                <cfStartDate>2010-01-05T00:00:00</cfStartDate>
        </cfResProd_Class>
</cfResProd>
```

Notes

Four dates supported in C4D for this element: temporal extent (which has a start and end date), created (single date), revised (single date), published (single date).

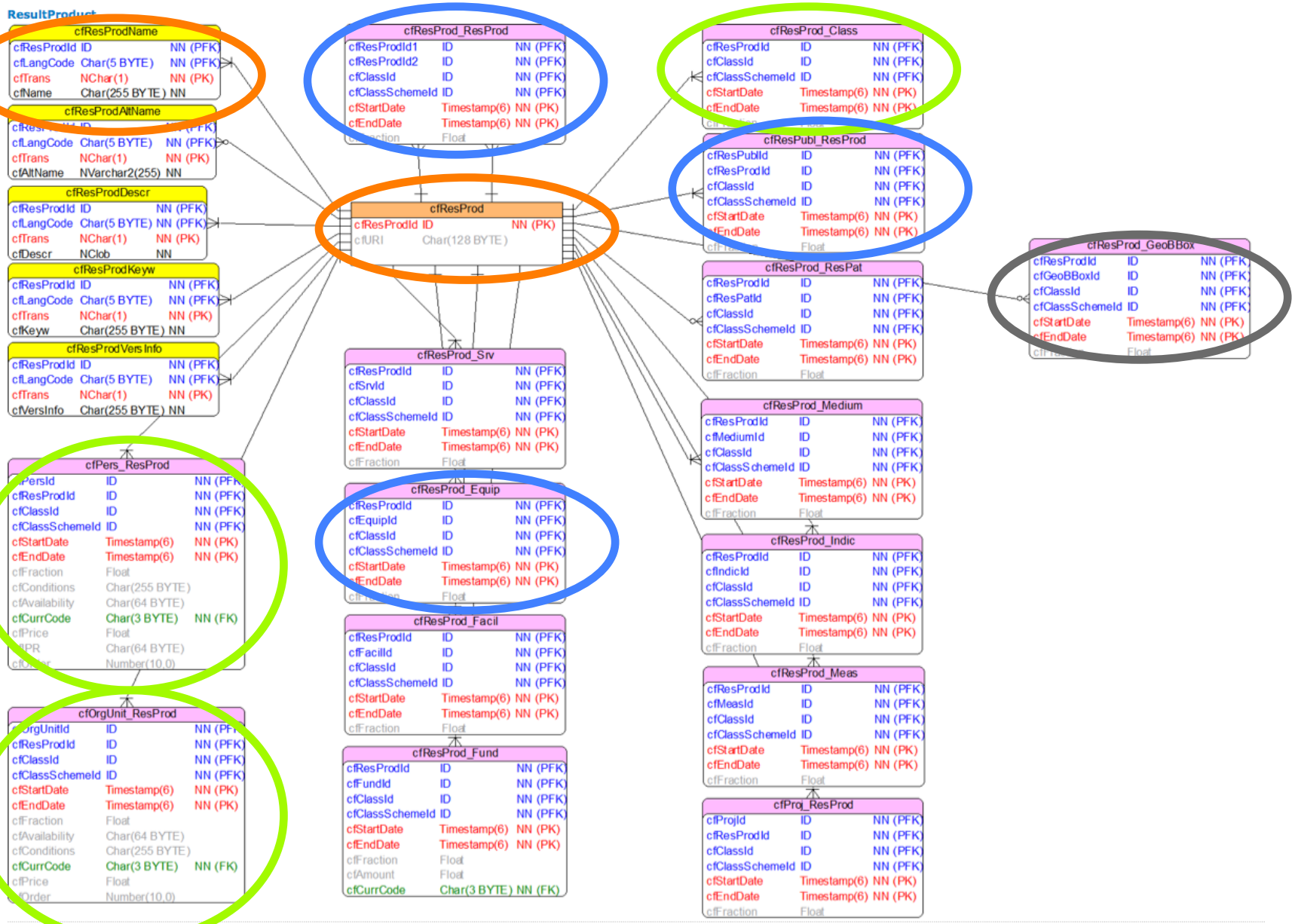
- cfClassId – one of: temporal_extent/publication/revision/creation
- cfClassSchemeId – always: class_scheme_resultProduct_classification_temporalReference

CERIF TG Notes

The CERIF TG suggests the temporal interval of the data itself (the effective datetime range of the observations) is of a different nature than the documentation of the dataset lifecycle. We would therefore suggest:

1. Expressing the temporal extent as two links to cfMeasurements that hold the startDateTime / endDateTime of the temporal reference.
2. Putting the creation/revision/publication timestamps as cfStartDates on the links to the parties responsible for the respective steps in the dataset's lifecycle.

CERIF 1.6 ER cfResProd



Conclusions – what worked well

- Mapping to CERIF pretty straightforward – because it already contains all the entities we need and most of the relationships
- Involving CERIF-TG meant we could give and take ideas – very constructive
- Modelling at the business level first helped resolve questions such as 'should this be a classification or a relationship to a person or organisation'; this is best practice anyway for sustainability and flexibility; why model a person or an organisation as an attribute … rather than separate entities; this is a fundamental fault with DC and similar 'flat' structures
- The separate "semantic layer" ie the classification schemes, allowed us to map different schemes (inspire themes, rcuk subject classifications for keywords) and seavox gazetteer for Extent (ie which bit of 'water')

Conclusions – more work needed

- Some MEDIN elements not fully modelled yet but tend to be full text fields so could be better to determine if can be broken down into more structured data, e.g Lineage (element 17)

- Agreeing semantics e.g. lifecycle stages of a dataset in order to properly model temporal aspects, such as published date, version date, created date, etc
- Translating conditions of (re)use into structural metadata (element 21); requires modelling at business level first

Actually, these aren't really issues with CERIF, more on business modelling and agreement on semantics and vocabularies …. Irrespective of data format.

Many thanks for listening

Anna Clements, Head of Research Data and Information Services
akc@andrews.ac.uk @annakclements

C4D Blog at http://cerif4datasets.wordpress.com/