

# CERIF COURSE

## Session 6: Evolution

Keith G Jeffery,

Director, IT CLRC

[k.g.jeffery@rl.ac.uk](mailto:k.g.jeffery@rl.ac.uk)

Anne Asserson,

University of Bergen

[anne.asserson@ub.uib.no](mailto:anne.asserson@ub.uib.no)

- How to propose changes
- Proposed changes
  - Dublin Core
  - Results/publications
  - Beat's classification scheme
- How to go forward

- How to propose changes
- Proposed changes
  - Dublin Core
  - Results/publications
  - Beat's classification scheme
- How to go forward

# Why Changes?

- CERIF is (we believe) a good model ★
- But it is not perfect ★
  - There may be errors or inconsistencies – especially in the least-used parts of the model e.g. classifications, enumerated lists of values
  - The end-user requirements change and evolve requiring new data structures to support them

## How to propose changes Changes to the model

---

- Changes (to the production model)
- Extensions (that may be accepted to the production model)
- Suggestions (that may be accepted to the production model)

- Changes are required when
  - An error has been found in the (updated) CERIF 2002 datamodel specification
  - There is a need to change the datamodel due to some external event e.g. legal change

- Extensions are improvements to the (updated) CERIF 2002 datamodel reflecting commonly required additional features. There are two kinds:
  - extension that does not affect core model
  - extension that does affect core model

## How to propose changes Change Process

- The change process is essentially the same
  - For changes to the CERIF model
    - Intended to correct an error or inconsistency
  - For extensions to the CERIF model
    - Intended to extend the model to meet a new user requirement

## How to propose changes Change Process

- Member proposes change to CERIF TG leader as follows
  - Rationale
  - Current CERIF model fragment (E-R and schema)
  - Proposed CERIF model fragment (E-R and schema)
  - URI to test change on proposer's test CERIF DB
- CERIF TG leader
  - validates at first level,
  - assigns Change id and sets up discussion forum thread
  - initiates process of asking TG members to handle

# How to propose changes Handle proposed change in TG

- CERIF TG members
  - Consider requirement and attempt alternative solutions without changing CERIF
  - If alternative works; CERIF TG leader asks TG for vote
  - If vote positive proposes to members for comment
  - Board endorses
- document and add whole documentation to best practice DB
  - If no alternative, then test change at URI
  - If proposal OK CERIF TG leader asks TG for vote
  - If vote positive proposes to members for comment
  - Board endorses
  - CERIF TG leader places changed model on website
- document and add whole documentation to best practice DB

- Metadata compatible
- CERIF 2002 compatible
  - How to certify?
  - Similar to change
  - Submission of E-R diagram and schema
  - CERIF TG leader initiates process like change
  - If OK get certification quality icon

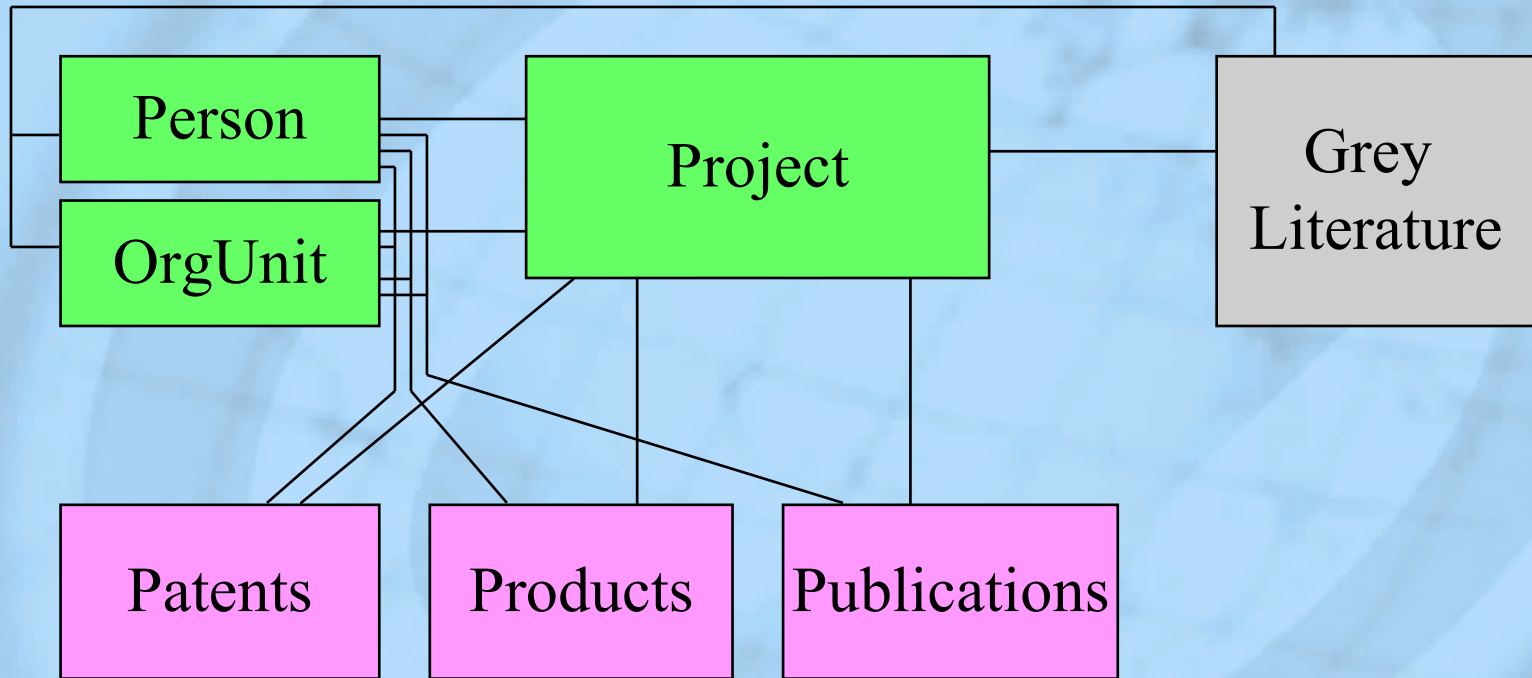
- extension that does not affect core model
  - Core model is the metadatamodel without the 'X\_Additional' tables
  - i.e. 3 primary base entities plus minimal language field base entities plus linking tables required
- extension that does affect core model

- How to propose changes
- Proposed changes
  - Dublin Core
  - Results/publications
  - Beat's classification scheme
- How to go forward

- How to propose changes
- Proposed changes
  - Dublin Core
  - Results/publications
  - Beat's classification scheme
- How to go forward

- The work started because of a perceived requirement for CERIF to handle Grey Literature
- (CERIF handles white literature by reference to external bibliographic databases)

# Adding DC to CERIF: Grey Literature



# Adding DC to CERIF: All literature

- Find that external references to white literature also insufficient
  - Because not all publications in external bibliographic databases
  - Because of need for additional information about the publication for various applications e.g. assignation to university department
  - But also that DC helps but not enough (see Publications proposed change)

- DC is a metadata standard W3C-approved
- very commonly used to describe e-resources on the web - not just publications but also datasets, programs and many multimedia items
- DC is machine readable but not machine understandable
- comes from the e-library area
- It consists of 15 elements (originally 13) and for each element there may be a scheme which defines an enumerated list of acceptable values.

## Purpose of the proposal

---

- define a formal version of Dublin Core which intersects where appropriate with CERIF entities and attributes
- be able to generate normal DC from this on demand for interoperation with other DC-metadata-compliant systems
- hopefully persuade the DC community that they need to move towards a much more formal DC to allow machine understanding in a GRIDs environment

- to be able to describe results of R&D stored in a CRIS (within CERIF)
  - Result\_Publication,
  - Result\_Patent,
  - Result\_Productin DC form

- Consists of 15 elements
- Each one seen as a <tag> or 'named section' of text
- The text may be controlled by allowed values (enumerated list)
- There are no relationships represented between <tags>

- **Element: Title**
- Name: Title
- Identifier: Title
- Definition: A name given to the resource.
- Comment: Typically, a Title will be a name by which the resource is formally known.
- CERIF provides language variants
- Not clear in DC if an unique ID of the resource or a descriptive name; CERIF has additional unique ID

- **Element: Creator**
- Name: Creator
- Identifier: Creator
- Definition: An entity primarily responsible for making the content of the resource. Comment: Examples of a Creator include a person, an organisation, or a service. Typically, the name of a Creator should be used to indicate the entity.
- CERIF: Person or OrgUnit, linked by Role and Time Period; repeating group may be required

- **Element: Subject**
- Name: Subject and Keywords
- Identifier: Subject
- Definition: The topic of the content of the resource.
- Comment: Typically, a Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.
- CERIF: split as one free text and other controlled terms. CERIF provides language variants

- **Element: Description**
- Name: Description
- Identifier: Description
- Definition: An account of the content of the resource.
- Comment: Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.
- CERIF provides language variants

- **Element: Publisher**
- Name: Publisher
- Identifier: Publisher
- Definition: An entity responsible for making the resource available
- Comment: Examples of a Publisher include a person, an organisation, or a service. Typically, the name of a Publisher should be used to indicate the entity.
- CERIF: OrgUnit or Person; repeating group may be required

- **Element: Contributor**
- Name: Contributor
- Identifier: Contributor
- Definition: An entity responsible for making contributions to the content of the resource.  
Comment: Examples of a Contributor include a person, an organisation, or a service. Typically, the name of a Contributor should be used to indicate the entity.
- CERIF: Person or OrgUnit; repeating group required

- **Element: Date**
- Name: Date
- Identifier: Date
- Definition: A date associated with an event in the life cycle of the resource.
- Comment: Typically, Date will be associated with the creation or availability of the resource.  
Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [[W3CDTF](#)] and follows the YYYY-MM-DD format.
- **CERIF: dates attached to linking relations i.e. provide a record of activity**

- **Element: Type**
- Name: Resource Type
- Identifier: Type
- Definition: The nature or genre of the content of the resource.
- Comment: Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the working draft list of Dublin Core Types [[DCT1](#)]). To describe the physical or digital manifestation of the resource, use the FORMAT element.

- **Element: Format**

- Name: Format
- Identifier: Format
- Definition: The physical or digital manifestation of the resource.
- Comment: Typically, Format may include the media-type or dimensions of the resource. Format may be used to determine the software, hardware or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types [[MIME](#)] defining computer media formats).

- **Element: Identifier**
- Name: Resource Identifier
- Identifier: Identifier
- Definition: An unambiguous reference to the resource within a given context.
- Comment: Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. Example formal identification systems include the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN).

- **Element: Source**
- Name: Source
- Identifier: Source
- Definition: A Reference to a resource from which the present resource is derived. Comment: The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.
- **CERIF: an identified source referenced through a link relation with appropriate role value and time period**

- **Element: Language**
- Name: Language
- Identifier: Language
- Definition: A language of the intellectual content of the resource.
- Comment: Recommended best practice for the values of the Language element is defined by RFC 1766 [[RFC1766](#)] which includes a two-letter Language Code (taken from the ISO 639 standard [[ISO639](#)]), followed optionally, by a two-letter Country Code (taken from the ISO 3166 standard [[ISO3166](#)]). For example, 'en' for English, 'fr' for French, or 'en-uk' for English used in the United Kingdom.
- **CERIF: handles multiple language variants**

- **Element: Relation**
- Name: Relation
- Identifier: Relation
- Definition: A reference to a related resource.
- Comment: Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.
- CERIF: linking relation with role and time period

- **Element: Coverage**

- Name: Coverage
- Identifier: Coverage
- Definition: The extent or scope of the content of the resource.  
Comment: Coverage will typically include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names [TGN]) and that, where appropriate, named places or time periods be used in preference to numeric identifiers such as sets of coordinates or date ranges.
- CERIF: split into temporal and spatial

- **Element: Rights**
- Name: Rights Management
- Identifier: Rights
- Definition: Information about rights held in and over the resource.
- Comment: Typically, a Rights element will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. If the Rights element is absent, no assumptions can be made about the status of these and other rights with respect to the resource.
- CERIF: handled as restrictive associative metadata: security, rights, privacy, pricing

- The principles used were:
  - Use existing CERIF Person, Project, OrgUnit – and link DC record(s) to these primary entities with roles
  - Use CERIF language-base technique to handle multilinguality correctly
  - Use CERIF linking relation technique to permit links and recursion
  - Add facility for annotation metadata

# DC-CERIF Intersection with CERIF

Person	PersonId	char(32)	m,pk		Creator (part)
Person	FamilyNames	char(48)	m	list, separated	Publisher (part)
Person	FirstNames	char(32)	o	list, separated	Contributor (part)
Person	OtherNames	char(32)	o	list, separated	
Person	Sex	char(1)	o,enumlist	M F ?	
Person	URI	char(128)		Person homepage	
OrgUnit	OrgUnitId	char(32)	m,pk		Creator (part)
OrgUnit	Acronym	char(16)	o		Publisher (part)
OrgUnit	Type	char(8)	m, enumlist	enumlist, separated	Contributor (part)
OrgUnit	Headcount	integer(8)	o	number of working staff	
OrgUnit	Turnover	float	o	total annual working budget	
OrgUnit	Currency	char(4)	m if value else o		
OrgUnit	URI	char(128)		OrgUnit homepage	
Project	ProjectId	char(32)	m,pk		Creator (part)
Project	StartDate	date	m		Publisher (part)
Project	EndDate	date	o		Contributor (part)
Project	Status	char(8)	m,enumlist	stalled, completed...	
Project	URI	char(128)		project homepage	

Entity	Attribute	Type	Constraint	Comments / Meaning	Compare DC
= CERIF					
<b>BASE TABLES</b>					
DC_Resource	DCId	char(32)	m,pk		Resource
DC_Resource	Scheme	char(32)	m,pk(part)	enumlist; see DC for valid schemes	
DC_Resource	ResourceId	char(128)		URI of the target being described	
DC_Resource_Type	DCId	char(32)	m,pk(part)		
DC_Resource_Type	Scheme	char(32)	m,pk(part)	enumlist; see DC for valid schemes	Resource Type
DC_Resource_Type	Resource_Type	char(16)		enumlist	
DC_Format	DCId	char(32)	m,pk(part)		Format
DC_Format	Scheme	char(32)	m,pk(part)	enumlist; see DC for valid schemes	
DC_Format	FormatType	char(8)		MIME	
DC_Format	Size	numeric		units from Scheme	

DC_Coverage_Spatial	DCId	char(32)	m,pk		Coverage (part)
DC_Coverage_Spatial	Scheme	char(16)	m,pk(part)	enumlist, includes units and projection	
DC_Coverage_Spatial	X-Coordinate	numeric			
DC_Coverage_Spatial	Y-Coordinate	numeric			
DC_Coverage_Spatial	Z-Coordinate	numeric			
DC_Coverage_Spatial	Precision	numeric			
DC_Coverage_Temporal	DCId	char(32)	m,pk		Coverage (part)
DC_Coverage_Temporal	Scheme	char(16)	m,pk(part)	enumlist; see DC for valid schemes	
DC_Coverage_Temporal	StartDateTime	date			
DC_Coverage_Temporal	EndDateTime	date			
DC_Coverage_Temporal	Precision	numeric			

DC_Rights_Management_Security	DCId	char(32)	m,pk		Rights Management (p
DC_Rights_Management_Security	Scheme	char(16)	m,pk(part)	enumlist	
DC_Rights_Management_Security	SecurityConstraint	char(64)			
DC_Rights_Management_Rights	DCId	char(32)	m,pk		Rights Management (p
Rights_Management_Rights	Scheme	char(16)	m,pk(part)	enumlist	
DC_Rights_Management_Rights	RightsConstraint	char(64)			
DC_Rights_Management_Privacy	DCId	char(32)	m,pk		Rights Management (p
DC_Rights_Management_Privacy	Scheme	char(16)	m,pk(part)	enumlist	
DC_Rights_Management_Privacy	PrivacyConstraint	char(64)			
DC_Rights_Management_Pricing	DCId	char(32)	m,pk		Rights Management (p
DC_Rights_Management_Pricing	Scheme	char(16)	m,pk(part)	enumlist	
DC_Rights_Management_Pricing	PriceConstraint	char(64)			

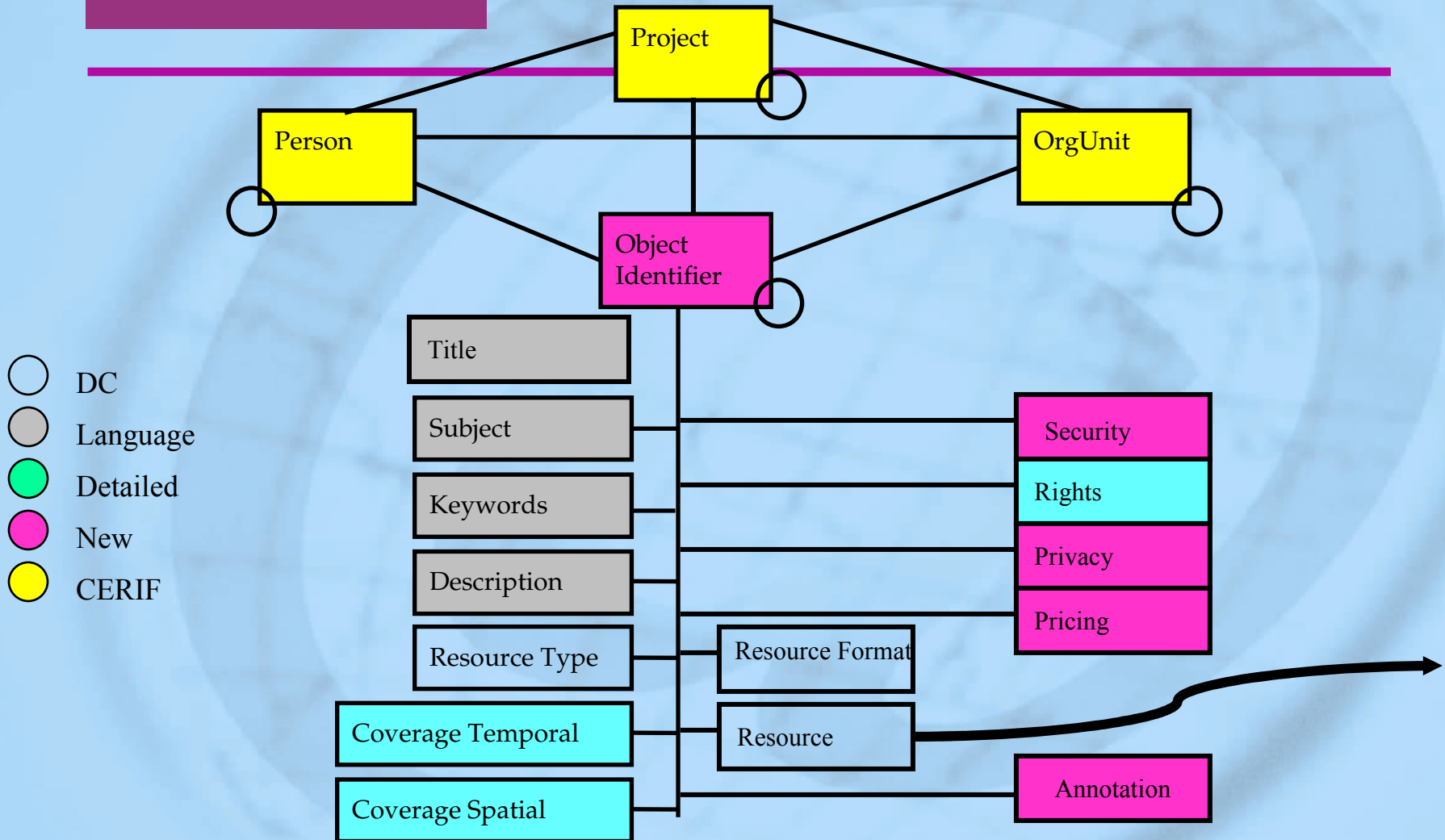
LANGUAGE BASE					
DC-Title	DCId	char(32)	m,pk(part)		Title
DC-Title	Scheme	char(16)	m,pk(part)	enumlist; see DC for valid schemes	
DC-Title	Language	char(2)	m,pk(part)		
DC-Title	Translation	char(1)	m,pk(part)	o(rig),h(uman),m(achine)	
DC-Title	Title	char(1024)			
DC-Subject	DCId	char(32)	m,pk(part)		
DC-Subject	Scheme	char(16)	m,pk(part)	enumlist; see DC for valid schemes	Subject (Part)
DC-Subject	Language	char(2)	m,pk(part)		
DC-Subject	Translation	char(1)	m,pk(part)	o(rig),h(uman),m(achine)	
DC-Subject	Subject	char(64)			
DC-Keywords	DCId	char(32)	m,pk(part)		Subject (Part)
DC-Keywords	Scheme	char(16)	m,pk(part)	enumlist; see DC for valid schemes	
DC-Keywords	Language	char(2)	m,pk(part)		
DC-Keywords	Translation	char(1)	m,pk(part)	o(rig),h(uman),m(achine)	
DC-Keywords	Keywords	char(1024)		comma separated list	

# DC-CERIF Language Base 2

DC-Description	DCId	char(32)	m,pk(part)		Description
DC-Description	Scheme	char(16)	m,pk(part)	enumlist; see DC for valid schemes	
DC-Description	Language	char(2)	m,pk(part)		
DC-Description	Translation	char(1)	m,pk(part)	o(rig),h(uman),m(achine)	
DC-Description	Description	char(3990)			
DC-Annotation	DCId	char(32)	m,pk(part)		*** new ***
DC-Annotation	Scheme	char(16)	m,pk(part)	enumlist	
DC-Annotation	Language	char(2)	m,pk(part)		
DC-Annotation	Translation	char(1)	m,pk(part)	o(rig),h(uman),m(achine)	
DC-Annotation	Annotation	char(3990)			

LINK TABLES				
DC-DC	DCId	char(32)	m, fk, pk(part)	Source
DC-DC	DCId	char(32)	m, fk, pk(part)	Relation
DC-DC	Role	char(16)	o, fk, pk(part), enumlist	
DC-DC	StartDate	date	o, fk, pk(part)	
DC-DC	EndDate	date	o, fk, pk(part)	
Person-DC	PersonId	char(32)	m, fk, pk(part)	Date
Person-DC	DCId	char(32)	m, fk, pk(part)	
Person-DC	Role	char(16)	o, fk, pk(part) e.g. author, editor, reviewer	
Person-DC	StartDate	date	o, fk, pk(part)	
Person-DC	EndDate	date	o, fk, pk(part)	
OrgUnit-DC	OrgUnitId	char(32)	m, fk, pk(part)	Date
OrgUnit-DC	DCId	char(32)	m, fk, pk(part)	
OrgUnit-DC	Role	char(16)	o, fk, pk(part) e.g. techreportediting, publishing	
OrgUnit-DC	StartDate	date	o, fk, pk(part)	
OrgUnit-DC	EndDate	date	o, fk, pk(part)	
Project-DC	ProjectId	char(32)	m, fk, pk(part)	Date
Project-DC	DCId	char(32)	m, fk, pk(part)	
Project-DC	Role	char(16)	o, fk, pk(part) e.g. techreportediting, publishing	
Project-DC	StartDate	date	o, fk, pk(part)	
Project-DC	EndDate	date	o, fk, pk(part)	

# CERIF-DC Data Model



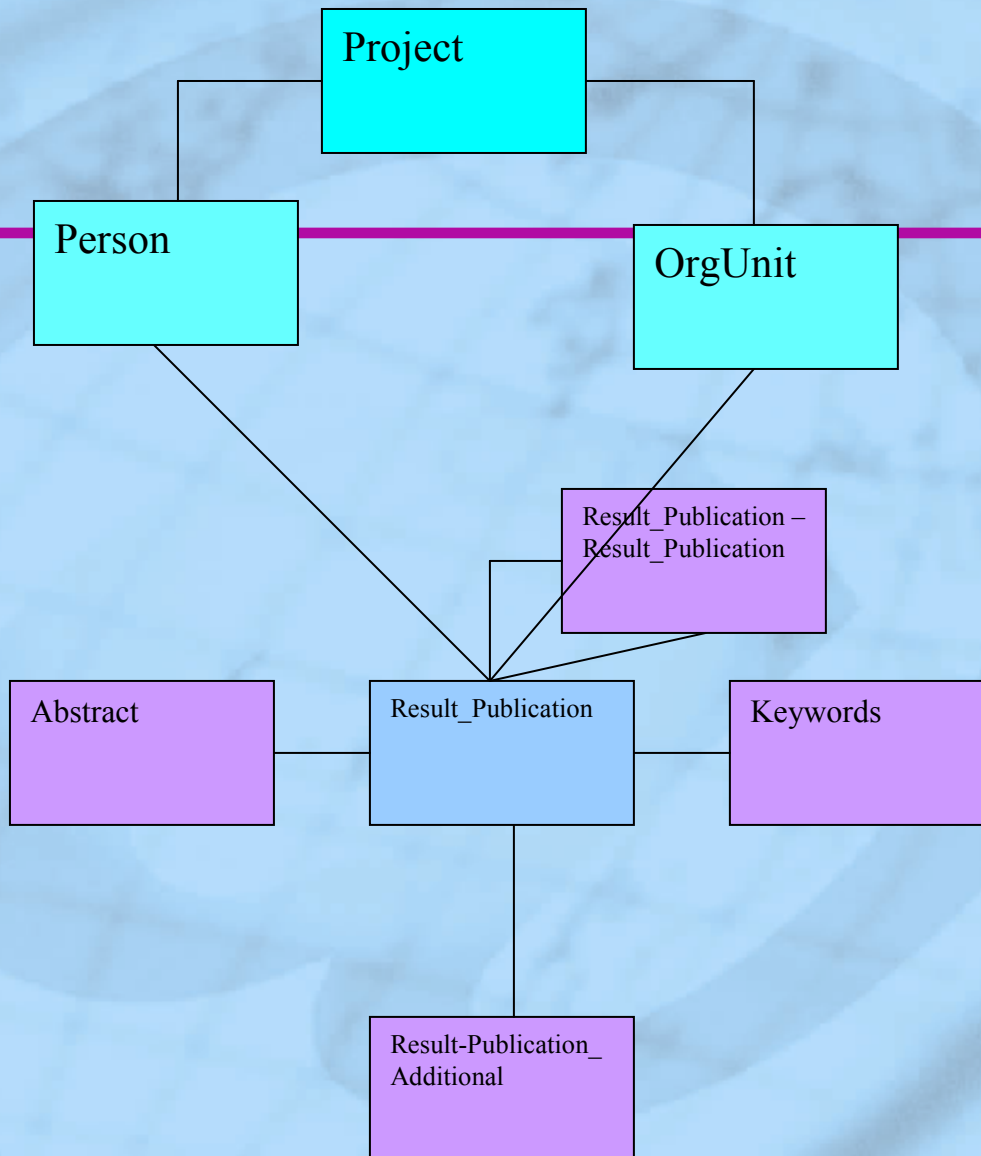
- Extensions in the data model to support Dublin Core

<http://www.eurocris.org>

About / Task Groups / CERIF / Proposed Changes / Extensions

- How to propose changes
- Proposed changes
  - Dublin Core
  - Results/publications
  - Beat's classification scheme
- How to go forward

- Will give the possibility for a better documentation of publications.
  - not all publications in external bibliographic databases
  - need for additional information about the publication for various applications e.g. assignation to university department
  - DC helps but not enough



Proposed changes  
Results/publications  
Necessary additions

---

- 1 Secondary table  
Result\_Publication\_Additional
- 2 Language tables  
Result\_Publication\_Abstract  
Result\_Publication\_Keywords
- 1 link table  
recursion on Result\_Publication-  
Result\_Publication

## Extension: Result\_Publication: Result\_Publication\_Additional

Result_Publication_Additional	Result_publicationll	char(32)	m
Result_Publication_Additional	Volume	char(4)	0
Result_Publication_Additional	Issue	char(4)	0
Result_Publication_Additional	From page	Char(4)	0
Result_Publication_Additional	To page	char(4)	0
Result_Publication_Additional	Total Pages	char(4)	0
Result_Publication_Additional	ISSN	char(16)	0,enumlist
Result_Publication_Additional	ISBN	char(16)	0,enumlist
Result_Publication_Additional	Refereed	char(1)	m; Y   N

## Extension: Result\_Publication: Abstract and Keywords

Result_Publication-Abstract	Result_PublicationKey	char(32)	m,pk(part)
Result_Publication-Abstract	Language	char(2)	m,pk(part)
Result_Publication-Abstract	Translation	char(1)	m,pk(part)
Result_Publication-Abstract	Abstract	char(3990)	
Result_Publication-Keywords	Result_PublicationKey	char(32)	m,pk(part)
Result_Publication-Keywords	Language	char(2)	m,pk(part)
Result_Publication-Keywords	Translation	char(1)	m,pk(part)
Result_Publication-Keywords	Keyword	char(1024)	

## Extension: Result\_Publication: Link Table

Result_Publication-Result_Pu	Result_PublicationI	char(32)	m, fk, pk(part)
Result_Publication-Result_Pu	Result_PublicationI	char(32)	m, fk, pk(part)
Result_Publication-Result_Pu	Role	char(16)	o, fk, pk(part), enumli
Result_Publication-Result_Pu	StartDate	date	o, fk, pk(part)
Result_Publication-Result_Pu	EndDate	date	o, fk, pk(part)

- Full data model (spreadsheet)

<http://www.ub.uib.no/avdeling/fdok/cris/taskgroups/FullDataModelkgjaa19980619rev200209201.xls>

- Search Fdok

<http://www.ub.uib.no/fdok/sok/>

- Task group webpage

<http://www.eurocris.org/taskgroups/cerif/index6.htm>

- How to propose changes
- Proposed changes
  - Dublin Core
  - Results/publications
  - Beat's classification scheme
- How to go forward

## Beat's classification scheme Based on

- CERIF Discipline Classification Schema, 1991
- Ortelius Thesaurus on Higher Education, 1988
- UNESCO International Standard Classification of Education ISCED, 1997
- Swiss University Information System, Technical Handbook, 2001  
Swiss National Science Foundation and ProClim Classification, 1996

- Uneven distribution
- Sub-disciplines and specializations in History, Linguistics, Law and Medicine are over-represented
- Leads to huge pick lists which are not satisfactory from an ergonomic point of view

# Beat's classification scheme

## Guiding principles of the present proposal

- Based on the CERIF 1991 classification, currently in use.
- The current discipline schema is compatible with the amendments proposed
- Consequent reduction of the data set.
- It thus aims to give a generative topical orientation and not a precise definition of (sub-) disciplines.

## Beat's classification scheme Guiding principles of the present proposal

---

- The result is a shortlist with two levels, which may be extended to three levels if required.
- The latter would consist of the currently used schema.
- The extended third level is, however, not revised in the present proposal.
- Its revision and consolidation would need a broad consent among European partners (i.e. governments, funding agencies, universities).

## Beat's classification scheme CERIF91 classification

Domain	Top areas		Disciplines	
	CERIF91	CERIF02	CERIF91	CERIF02
Humanities	4	18	87	123
Social Sciences	6	21	76	107
Physical Sciences	8	21	55	87
Biomedical Sciences	10	30	91	152
Technological Sciences	6	23	60	81

- Beat Sottas and his team
- <http://www.aramis-research.ch/e/ProposalClassification2002-10.xls>

- How to propose changes
- Proposed changes
- Dublin Core
- Results/publications
- Beat's classification scheme
- How to go forward

- euroCRIS CERIF Task Group has the responsibility from the EC to maintain and develop
  - CERIF
  - Expertise to assist people to use CERIF
- CERIF is a living, evolving datamodel aimed at serving the CRIS community
- You are invited to join in and help!