Strasbourg IHU Knowledge base: a CERIF implementation

Table of contents

Introduction

Nowadays, aggregating data seems the easiest thing to do. Big data is a popular concept that everyone wants to be part of. Connected objects flood the Internet with huge amounts of data. Even in research organizations, each person aggregates data at her/his own personal level. Complexity comes with the categorization of this amount of knowledge, at this level. For an organisation, complexity can also be increased when talking about sharing this knowledge among people or services.

Strasbourg IHU is an organisation which field of activity concerns image-guided minimally invasive surgery. This organisation is involved in innovation through partnerships with companies in their field. Many people in this organisation aggregate and produce data. They found they have problems capitalising on these data that are poorly shared across the organisation. So they started a project to create a platform to help knowledge sharing, and enhance this knowledge by getting relevant news concerning their specific fields.

This document will first give us an overview of the stakeholders and a description of the objectives of the project, and of the different steps of the analysis. A focus will then be done on the use of CERIF in the implementation of the platform, for one of its module. A third part will talk about the two other modules that are part of the platform, and we will close the document with some perspective for the future.

I) Project scope

1) Stakeholders

Strasbourg IHU[1]

Strasbourg IHU (Institut Hospitalo-Universitaire, or teaching hospital institute) is a unique medical and surgical centre dedicated to the management of digestive diseases. It combines the best minimally invasive technologies with the latest advances in medical imaging.

Its clinical activity is aligned with the one of the medical and surgical hepatico-digestive pole of the new civil hospital of Strasbourg, with a focus on the treatment of cancers of the digestive system.

The convergence of gastrointestinal surgery, flexible endoscopy and intraoperative imaging is a strategic priority. A major goal is to integrate the resources of surgical robotics and medical imaging to a new concept of hybrid surgery room.

With the status of Scientific Cooperation Foundation, IHU brings together public and private partners to develop, in the field of image-guided minimally invasive surgery, a program of excellence in care, training, search and recovery.

IS4RI[2]

IS4RI stand for Information Systems for Research and Innovation. The goal of this structure is to support the design, development, integration and use of information systems to contribute to the advancement of research and innovation in Europe and beyond.

It stands on 4 pillars:

- Cooperation: participate to the defragmentation of European research
- Valorisation: obtain more transparency, diffusion and reuse of research results
- Education: develop a European profession of research information manager
- Evaluation: assist in measuring research performance

2) Project description

IHU is a worldwide reference for image-guided minimally invasive surgery. As such, it interacts with many specialists including leader experts. Through these interactions and the research programs it conducts, IHU produces and aggregates amounts of advanced information about the field of minimally invasive surgery. This information concerns the whole technology development process, from invention to distribution. This knowledge is a real value for IHU, which wants to leverage it, especially through the creation of a dedicated digital service, designed as a platform composed of three modules.

The first module is described as a knowledge base. It allows users to add structured data and link them together in a database. All data inserted within the knowledge base should be categorised using a specific thesaurus. A search engine is used to search through the database. Users can navigate through data using faceted search, based on the categories described in the thesaurus.

[1] http://www.ihu-strasbourg.eu
[2] http://www.is4ri.com/

The second module is an automated market watch module. Its aim is to harvest information about image-guided minimally invasive surgery in different sources across the Internet. Information coming from this harvesting engine must then be sorted and qualified by IHU experts.

The third module concerns social network features. This module aims to ease interactions with experts, companies and medtech stakeholders, in the specific field of minimally invasive surgery. This module lets users comment and share content from the knowledge base or the market watch, and includes several tools to exchange news and information.

3) Project analysis

The project has been conducted in three phases: analysis, proof of concept and implementation.

Target

First, an analysis has been led to identify profiles of the actors of the platform. The following profiles have been selected:

- Fellow surgeon (research and innovation)
- Senior surgeon
- Radio manipulator
- Research engineer
- Development engineer
- Veterinary
- External start-up
- Simulation researcher
- Health economist

A campaign of interviews has been conducted with all selected profiles to precise their needs related to the platform. Interviews revealed a lack of databases to pool data and documents. Information is kept by each user on her/his own computer. No tool exists within IHU to share documents and data and centralise them.

The second redundant item was the difficulty to identify the right information holder. Employees do not know how to identify a potential contact about a specific field of investigation. This item does not talk about expertise, but about information available within the IHU and the right person to ask for this information or for additional details concerning this information. For example, if an employee attends a conference and gets some information about a particular technology, this person could not be identified as the one to ask for information about this technology.

The third item that has to be taken into account is the lack of information about other teams within IHU, coupled to a lack of visibility of people's own work and needs. IHU employees wish they could share news and articles, or just be informed about current projects running in IHU.

An additional study has been conducted to get the most common information that may be searched by users. This study identified several types of content:

- Publications, mainly coming from the PubMed[1] repository
- Market and medico-economic information (companies, market studies)
- Technical medical data (procedures, pathologies)
- Intellectual Property data (patents)

[1] http://www.ncbi.nlm.nih.gov/pubmed

- Devices information (instruments)

Proof of concept

The types of content identified by the analysis have then been implemented in a proof-of-concept (PoC) platform. The project team identified the metadata than needed to be stored for each entity, leading to the scheme in figure 1.

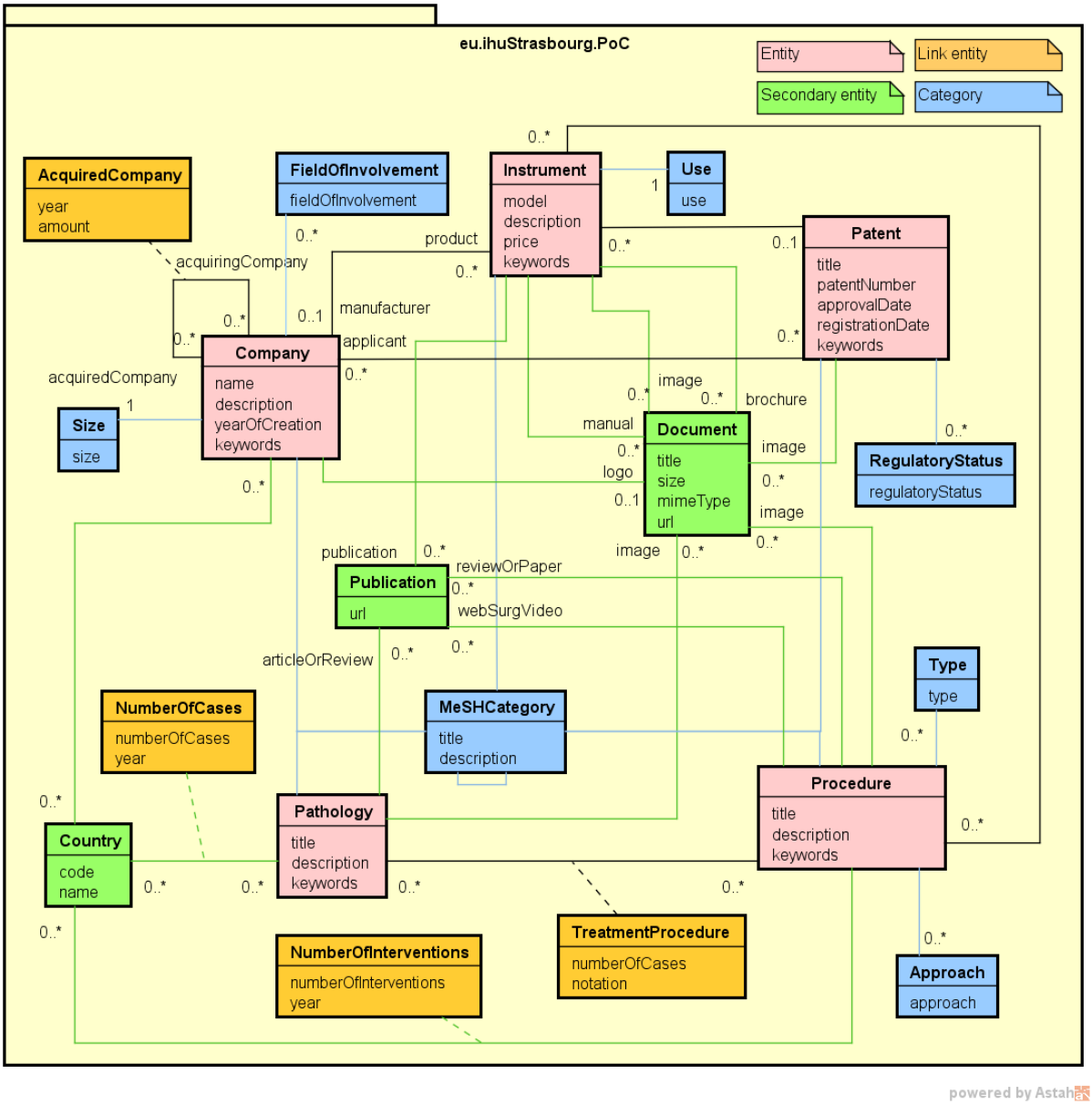


Figure 1: knowledge base scheme

These five main entities compose the first version of the database. Company, Patent and Instrument describe the medico-economic, IP and devices parts of the knowledge base. An instrument can be linked to a procedure which is used for some pathologies, describing the technical medical part. Publications are stored as links (URLs) to resources available on the Internet. All entities can be categorised using specific vocabularies (regulatory status, size, etc) and by using MeSH.

MeSH[1] stands for Medical Subject Headings. It is a vocabulary thesaurus used for indexing articles for PubMed, a repository for biomedical and life sciences publications. It is maintained by the US National Library of Medicine. The whole thesaurus is not interesting for the knowledge base. It seems that only four top categories should be kept for the platform: “Anatomy (A)”, “Diseases (C)”, “Analytical, Diagnostic and Therapeutic Techniques and Equipment (E)” and “Information Science (L)”. Use of the PoC helps the project team to determine if these categories are consistent to describe a set of representative content.

Technically, the PoC is implemented on a Liferay[2] platform. Liferay Portal is described as the leading open source portal for the enterprise. It offers content management, collaboration, and social features out-of-the-box. It also offers a wide API (Application Programming Interface) and is compliant to a large variety of standards, which ease integration and adaptability. The PoC lets the project team validate the choice of Liferay Portal, and check whether out-of-the-box functionalities fulfil the needs or if adaptations or new functionalities should be specifically developed.

The PoC uses Liferay’s content management for the knowledge base, with specific settings to store the five main entities. Liferay’s content management stores content as assets, which can be related to each other. Liferay also includes categories management that can be used to describe assets. These categories can then be used as facets in the internal search engine, with several other metadata, like author, dates, type of content, etc.

The market watch and collaboration modules were not included in the PoC. The market watch module had to be fully developed as it does not exist as a plugin or as an out-of-the-box functionality. The social part of the platform was considered as a secondary objective, and it has been decided to use basic Liferay features for a first version (rating and commenting assets).

The PoC has been tested and the knowledge base filled with data for several weeks by beta-users from IHU (fellow surgeons and research engineers). It showed that the selected MeSH categories were too wide to be used efficiently, and needed to be truncated for a good user experience. Liferay’s search engine and asset management do the job for the knowledge base, but the related-asset system does not provide enough details to describe the roles of each entity in a relation. Some specific development needs to be done to fulfil this need.

[1] https://www.nlm.nih.gov/mesh/meshhome.html
[2] https://www.liferay.com

II) Knowledge-base: working with CERIF

1) CERIF-structured database

With the need for specific development for the knowledge base, the project team had to find a structure to store data in the database.

CERIF is a data model that allows to store research information. Information is described as entities that contains specific metadata, for basic concepts (organisation units, project, etc.), research results (publications, patents and products), infrastructures (equipment, facilities, etc.) and several secondary entities. A full part of the model also allows to describe measurements.

Most of the entities needed for the knowledge base are already available in CERIF. The figure 2 describes the CERIF entities used for each kind of information needed. The only entity that cannot be linked to any CERIF entity is the pathology.

Knowledge base type of content	Corresponding CERIF entity
Publication	CfResultPublication
Company	CfOrganisationUnit
Procedure	CfResultProduct
Pathology	?
Patent	CfPatent
Instrument	CfEquipment

Figure 2: knowledge base type of content and corresponding CERIF entities

CERIF also provides description for the links between entities. These relation entities have been used to describe relations within the knowledge base.

The measurements part of CERIF has been used to store all amounts (number of cases for pathologies and number of interventions for procedures).

CERIF provides a full semantic layer that is used for two goals: to describe roles for all links entities, and to provide additional information for entities. The knowledge base takes advantage of these two uses of the semantic layer.

A full mapping of the knowledge base to the CERIF model can be found in appendix 1.

2) Extending CERIF

Pathology is a type of content that cannot be represented in CERIF. To complete the CERIF model to fulfil the knowledge base needs, an extension has been designed to integrate a new entity. This part of the model has been built using the standard CERIF model for entities and link-entities. The basic modelling of pathologies is displayed in figure 3.

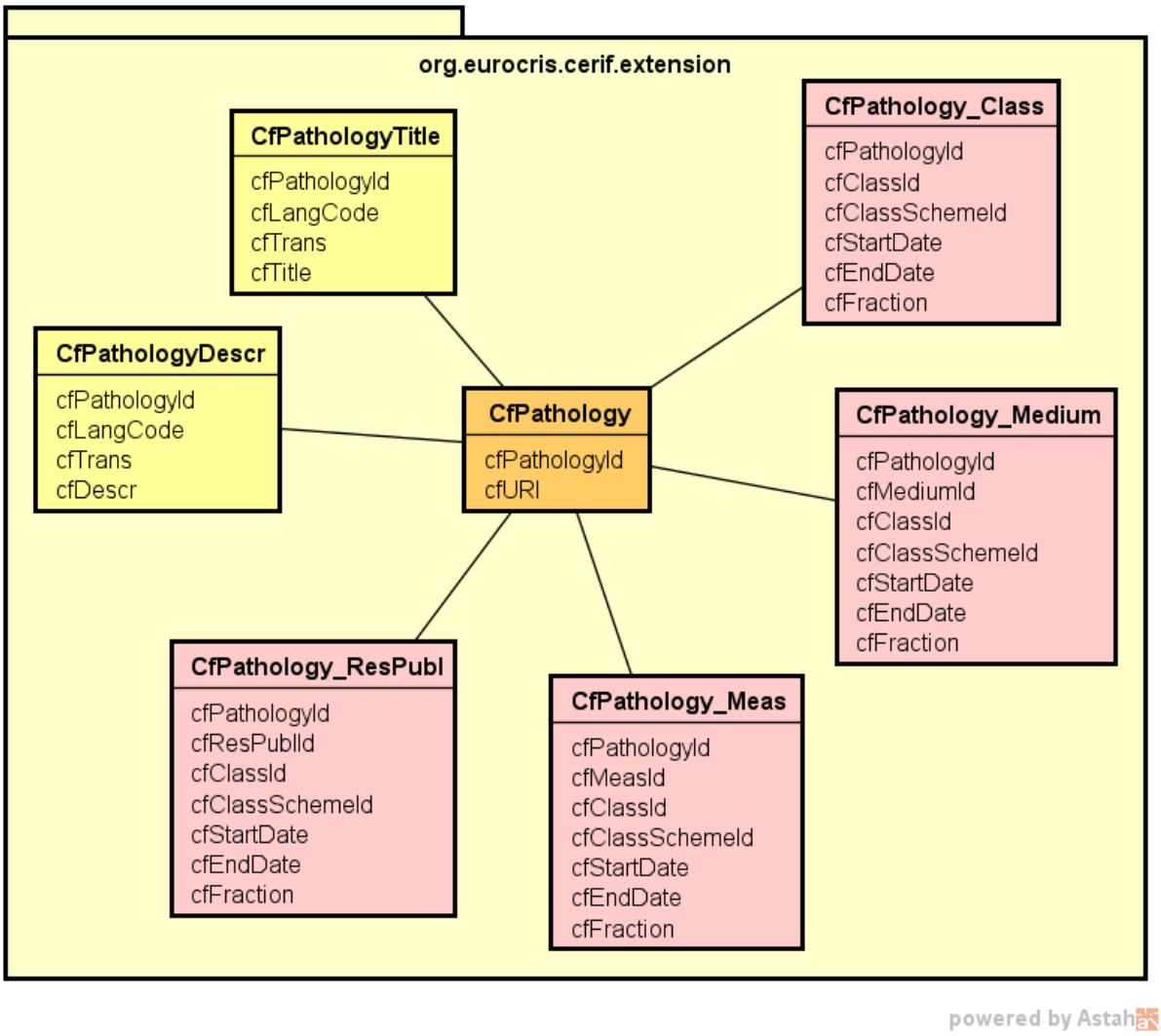


Figure 3: CERIF extension for pathology

The part of the model uses several principles designed in CERIF. The full text metadata for title and description uses the translation system to allow multiple languages to be used. Links with other entities include the semantic layer to describe their roles. They also implement the starting and ending dates to reflect evolution of this relation. The fraction completes the description of links to integrate the fact that a particular occurrence of a relation may only be a part of the real relation between two entities. The semantic layer is also used to complete the description of the pathology entity.

3) CERIF denormalisation

In the development process, some changes have been done to the normalised CERIF model. These changes mainly serve two goals.

First, this denormalisation improves integration within the Liferay Platform. Liferay comes with an API that eases development, giving developers a lot of tools to automatically generate fully-compatible code. This allows developers to create new types of assets that can take advantage of out-of-the-box features, thus saving them a huge amount of time. In this context, Liferay's categorisation has been used to complete the description of entities; CERIF semantic layer is only used to define roles for the relations. Liferay's tagging has also been used instead of CERIF keyword's metadata. Liferay

also includes a translation mechanism that has been used for all metadata that use CERIF multilingual system.

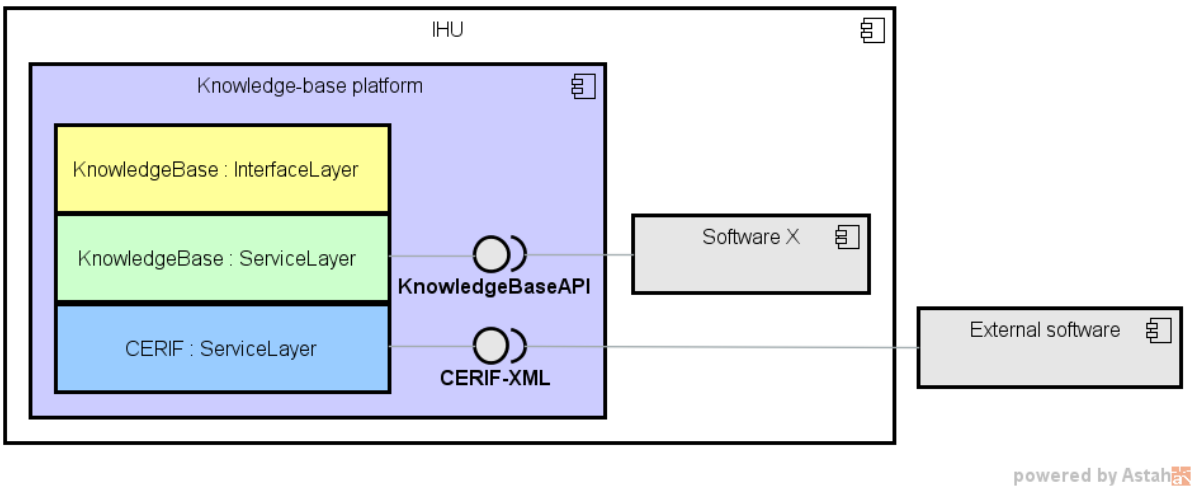
The second goal of the denormalisation is to ease the use of CERIF. In this context, the measurement part of the model has been simplified to use the semantic layer (cfClassification and cfClassificationScheme) to store indicators, instead of the dedicated entity cfIndicator. The reason of this choice is that cfIndicator does not provide more (nor different) metadata than cfClassification for the use of the knowledge base. CfIndicator mainly provides a way to identify what the value – stored in cfMeasurement – represents. This could be done using the cfClassificationId foreign key in cfMeasurement as well, as the knowledge base does not need to use this particular metadata for any other use.

CERIF denormalised diagrams are available in appendix 2.

4) Overlaying CERIF

CERIF is a meta-model that can be used for a wide variety of uses. Thus leading to a complex model that may lead to a complex interface to use it as-is.

To ease the look and feel of the software and to improve the user experience, multiple layers of service have been implemented. CERIF is used as the model to structure data in the database. A service layer has been implemented to access this data and manipulate CERIF entities. On top of this service, another service layer has been implemented that allows to manipulate knowledge base entities (company, instrument, patent, procedure and pathology, cf. figure 3). This second layer implements the logic that transforms knowledge base entities to CERIF entities, regarding the mapping (cf. appendix 1). This second service layer is then used by the interface layer to display information to the user.



This stack of services allows to manage mandatory fields that are not indicated as such in CERIF. It gives the user an interface that contains only the knowledge base metadata. It also gives the possibility to develop several API to access knowledge base information: either using the knowledge base layer (for IHU internal use for example) that could returns knowledge base entities, either using the CERIF layer to return CERIF-XML that will be understandable by any CERIF-compliant software.

III) Other modules implementation

1) Market watch

The goal of this module is to provide relevant news about a specific market (mini-invasive surgery), by harvesting and categorising content from defined sources. To reach this goal, several steps have been designed.

The first step currently implemented is to harvest periodically different content from several identified sources. These sources are specific to the domain (RSS feeds from domain-specific websites, PubMed repository, European Patent Office, etc.) or more generic, like search engines. Users can view and sort harvested content by sources, and can categorise content. Authorised users can also delete non-relevant content. This first step allows to determine which sources are the best, and build a corpus of categorised elements.

In addition to categorised content from the knowledge base, this corpus can then be used to calibrate a machine learning algorithm. This second step will allow further harvested items to be automatically categorised. The same principle can also be applied to the knowledge base to automatically propose categorisation for content inserted by users.

Categorised content from relevant sources is expected to help innovation and knowledge sharing within the targeted domain.

2) Social networking

The social networking module has been designed to promote knowledge sharing, and ease interactions between all actors in innovation: researchers, start-ups, etc.

The current module implements rating and commenting for any asset of the platform, including knowledge base entities and market watch items. These basic features are known to increase participation in a platform. Relevance of the harvested sources can also be defined using the rating as one of the criteria. Comments for an asset are starting points for interactions between different groups of users that potentially do not have interactions in their daily job.

The next step is the profiling of users. This feature lets users describe and enrich their profile, and search for specific skills so that they can increase their network to help solve specific problems or discover new fields. Profiling can also be enriched by harvesting several repositories using well-known identifiers like ORCID for example.

In addition to that, participation can be encouraged using gamification. Gamification is a principle that allows users to gain badges or trophies, depending on their participations in a platform. With a system of points, earned by writing content in the knowledge base, commenting and rating assets, or updating existing content, users will increase their level of experience, and display their badges in their profile, or be put forward as best contributors of the moment for example. This is a known method to increase participation.

Conclusion

Before giving users access to the platform, the platform has been integrated in the IHU IT environment. An IHU internal web designer worked on the interface, which is compliant with IHU graphic charter. The platform has also been integrated into the local IT infrastructure, with a close collaboration with the IHU IT team; it is compliant with their standards, like Single Sign On (SSO) for example.

In February 2016, the platform is open to users and the IHU part of the project team supports groups of users to gradually join and use the platform. This is a major phase of the project, to ensure a better adoption of the tool.

All three deployed modules give a first level of implementation, which can be improved. Suggestions have been collected and proposed for future versions. For the knowledge base, automatic indicators should be put in place to reveal its use. An extension has already been proposed in term of type of content, with entities like experts and projects. The harvesting engine should be refined by the analysis of the relevance of results and the refinement of keywords used and sources. A machine learning algorithm should be also be studied and calibrated so that content could be automatically categorised. Last but not least, improvements could also be made to the social network part, with user profiles, that could be automatically enriched with external sources.

The development of this platform, in this first phase, has set the foundations of a set of tools contributing to better knowledge discovery and management in the domain of minimally-invasive surgery. CERIF has proved useful to jump start the knowledge model definition, and combined to Liferay, it quickly provided for the necessary Proof-of-Concept that convinced the users.

Appendix

1) CERIF mapping for the knowledge base

Company

Knowledge base entity	Knowledge base attribute	CERIF entity	CERIF attribute	Comment
Company	name	cfOrgUnit	cfName	
Company	description	cfOrgUnit	description	
Company	yearOfCreation	cfOrgUnit	yearOfCreation	
Company	keywords			Liferay tags
Company	product	cfOrgUnit_Equip	cfEquipId	Class=manufacturer ClassScheme=Link between equipments and organisations
Company	logo	cfOrgUnit_Medium	cfMediumId	Class=logo ClassScheme=Link to medium
Company	patent	cfOrgUnit_ResPat	cfResPatId	Class=applicant ClassScheme=Link between patterns and organisations
Company	country	cfOrgUnit_PAddr cfPAddr	cfCountryCode	Class=location ClassScheme=Physical Address for Organisation Unit
Size	size			Liferay vocabulary
FieldOfInvolvement	fieldOfInvolvement			Liferay vocabulary
AcquiredCompany	acquiringCompany	cfOrgUnit_OrgUnit	cfOrgUnitId1	Class=acquired ClassScheme=Organisation Units relations
AcquiredCompany	acquiredCompany	cfOrgUnit_OrgUnit	cfOrgUnitId2	
AcquiredCompany	year	cfOrgUnit_OrgUnit	cfStartDate	
AcquiredCompany	amount	cfOrgUnit_OrgUnit	amount	

Instrument

Knowledge base entity	Knowledge base attribute	CERIF entity	CERIF attribute	Comment
Instrument	model	cfEquip	cfName	
Instrument	description	cfEquip	cfDescr	
Instrument	price	cfEquip	cfPrice	
Instrument	keywords			Liferay tags
Instrument	manufacturer	cfOrgUnit_Equip	cfOrgUnitId	Class=manufacturer ClassScheme=Link between equipments and organisations
Instrument	publication	cfResPubl_Equip	cfResPublId	Class=talk about ClassScheme= Publications about equipments
Instrument	image	cfEquip_Medium	cfMediumId	Class=image ClassScheme=Link to medium
Instrument	manual	cfEquip_Medium	cfMediumId	Class=manual ClassScheme=Link to medium
Instrument	brochure	cfEquip_Medium	cfMediumId	Class=brochure

				ClassScheme=Link to medium
Instrument	patent	cfResPat_Equip	cfResPatId	Class=protected by ClassScheme=Link between equipments and patents
Instrument	procedure			Not implemented yet
Use	use			Liferay vocabulary

Patent

Knowledge base entity	Knowledge base attribute	CERIF entity	CERIF attribute	Comment
Patent	title	cfResPat	cfTitle	
Patent	patentNumber	cfResPat	cfPatentNum	
Patent	approvalDate	cfResPat	cfApprovDate	
Patent	registrationDate	cfResPat	cfRegistrDate	
Patent	keywords			Liferay tags
Patent	applicant	cfOrgUnit_ResPat	cfOrgUnitId	Class=applicant ClassScheme=Link between patterns and organisations
Patent	instrument	cfResPat_Equip	cfEquipId	Class=protected by ClassScheme=Link between equipments and patents
Patent	image	cfResPat_Medium	cfMediumId	Class=image ClassScheme=Link to medium
RegulatoryStatus	regulatoryStatus			Liferay vocabulary

Procedure

Knowledge base entity	Knowledge base attribute	CERIF entity	CERIF attribute	Comment
Procedure	title	cfResProd	cfName	
Procedure	description	cfResProd	cfDescr	
Procedure	keywords			Liferay tags
Procedure	instrument			Not implemented yet
Procedure	image	cfResProd_Medium	cfMediumId	Class=image ClassScheme=Link to medium
Procedure	reviewOrPaper	cfResPubl_ResProd	cfEquipId	Class=talk about ClassScheme=Publications about pathologies
Procedure	webSurgVideo	cfResPubl_ResProd	cfEquipId	Class=websurg videos ClassScheme=Publications about pathologies
Type	type			Liferay vocabulary
Approach	approach			Liferay vocabulary

TreatmentProcedure	procedure	cfResProd_Meas	cfResProdId	Class= Treatment procedure measurement participation ClassScheme= Measurements
TreatmentProcedure	pathology	cfPathology_Meas	cfPathologyId	Class= Treatment procedure measurement participation ClassScheme= Measurements
TreatmentProcedure	numberOfCases	cfMeas	cfCountInt	
TreatmentProcedure	notation	cfMeas	cfValFloatP	
NumberOfInterventions	procedure	cfResProd_Meas	cfResProdId	Class= Procedure intervention measurement participation ClassScheme= Measurements
NumberOfInterventions	country	cfCountry_Meas	cfCountryCode	Class=Procedure intervention measurement participation ClassScheme= Measurements
NumberOfInterventions	numberOfInterventions	cfMeas	cfCountInt	
NumberOfInterventions	year	cfMeas	cfDateTime	

Pathology

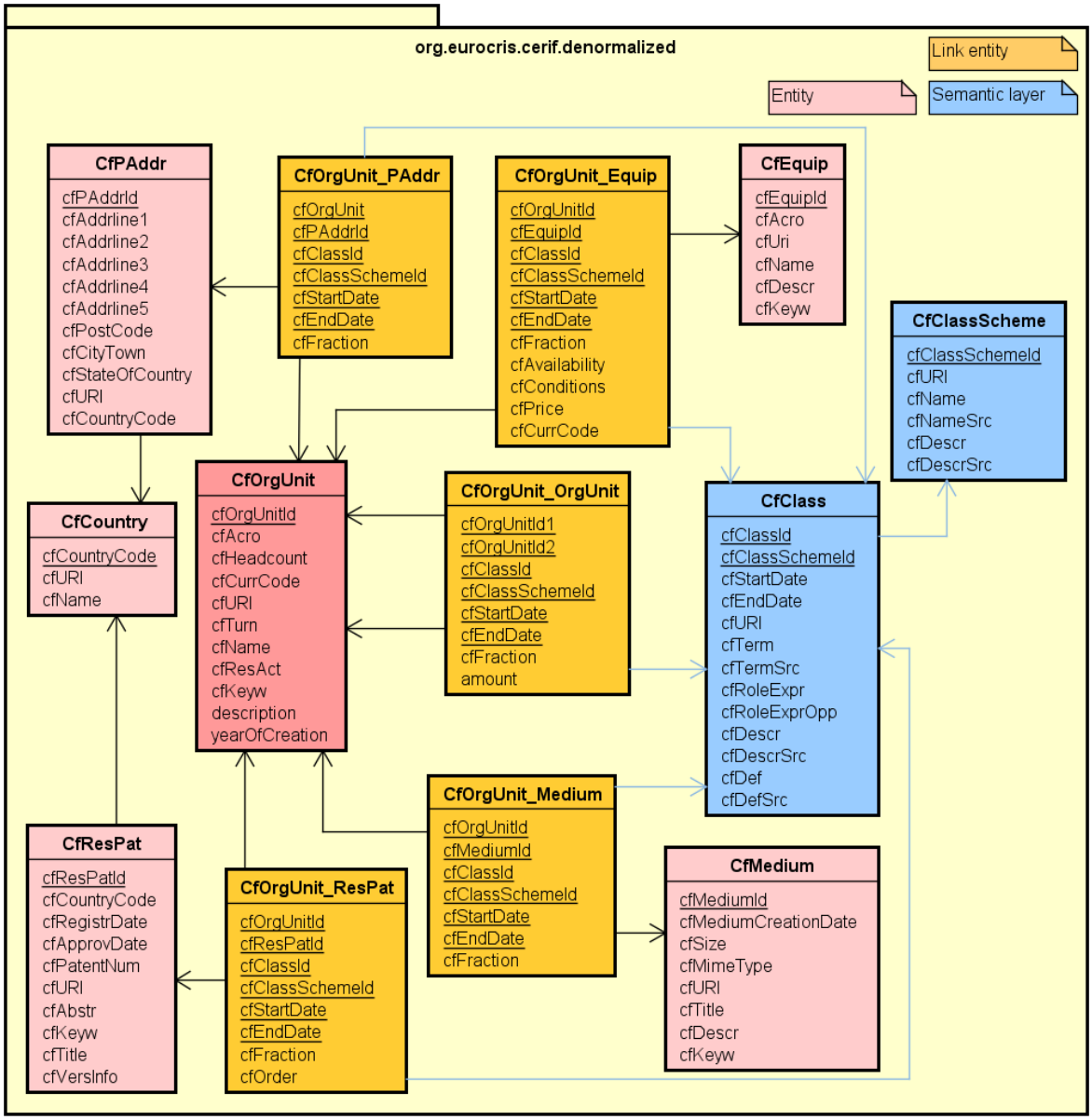
Knowledge base entity	Knowledge base attribute	CERIF entity	CERIF attribute	Comment
Pathology	title	cfPathology	cfTitle	
Pathology	description	cfPathology	cfDescr	
Pathology	keywords			Liferay tags
Pathology	image	cfPathology_Medium	cfMediumId	Class=image ClassScheme=Link to medium
Pathology	articleOrReview	cfPathology_ResPubl	cfPathologyId	Class=talk about ClassScheme=Publications about pathologies
TreatmentProcedure	procedure	cfResProd_Meas	cfResProdId	Class=Treatment procedure measurement participation ClassScheme= Measurements
TreatmentProcedure	pathology	cfPathology_Meas	cfPathologyId	Class=Treatment procedure measurement participationClassScheme = Measurements

TreatmentProcedure	numberOfCases	cfMeas	cfCountInt	
TreatmentProcedure	notation	cfMeas	cfValFloatP	
NumberOfCases	pathology	cfPathology_Meas	cfPathologyId	Class=Pathology case measurements participation ClassScheme= Measurements
NumberOfCases	country	cfCountry_Meas	cfCountryCode	Class=Pathology case measurement participation ClassScheme= Measurements
NumberOfCases	numberOfCases	cfMeas	cfCountInt	
NumberOfCases	year	cfMeas	cfDateTime	

2) Denormalised CERIF diagrams

CERIF entities used for company

The main entity to represent companies in CERIF is cfOrgUnit.

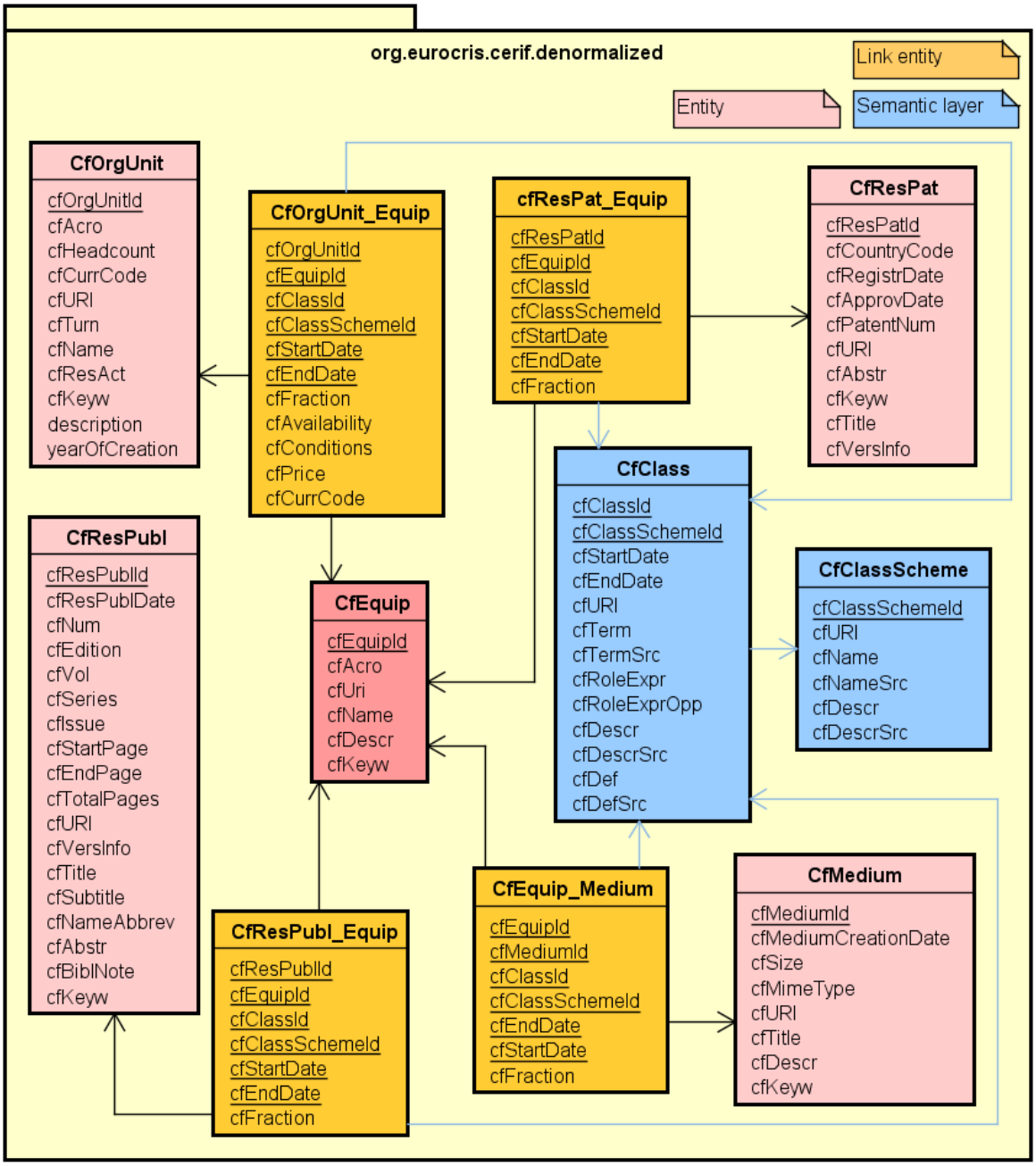


A few additional metadata have been added to the model:

- cfOrgUnit.description: this metadata allows users to describe a company
- cfOrgUnit.yearOfCreation: this metadata is used to store the creation date of the company; this metadata is a misuse of the CERIF model: it should have been stored as a measurement for the corresponding cfOrgUnit
- cfOrgUnit_OrgUnit.amount: this metadata is used to store the acquisition amount of an organisation unit; this metadata is also a misuse of the CERIF model: it should have been stored as a measurement for the corresponding cfOrgUnits (acquired and acquiring companies)

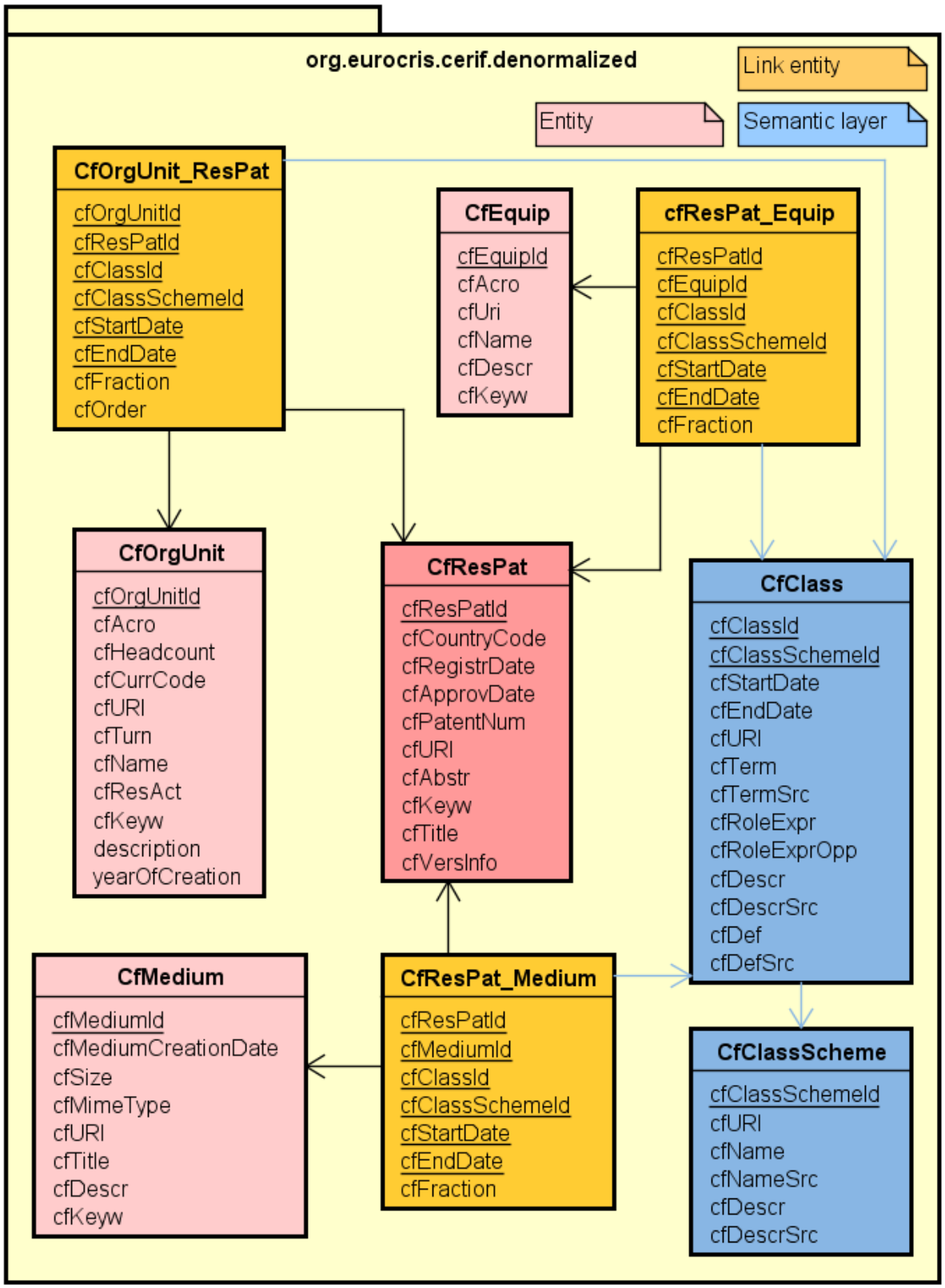
CERIF entities used for instrument

The main entity to represent instruments in CERIF is cfEquip.



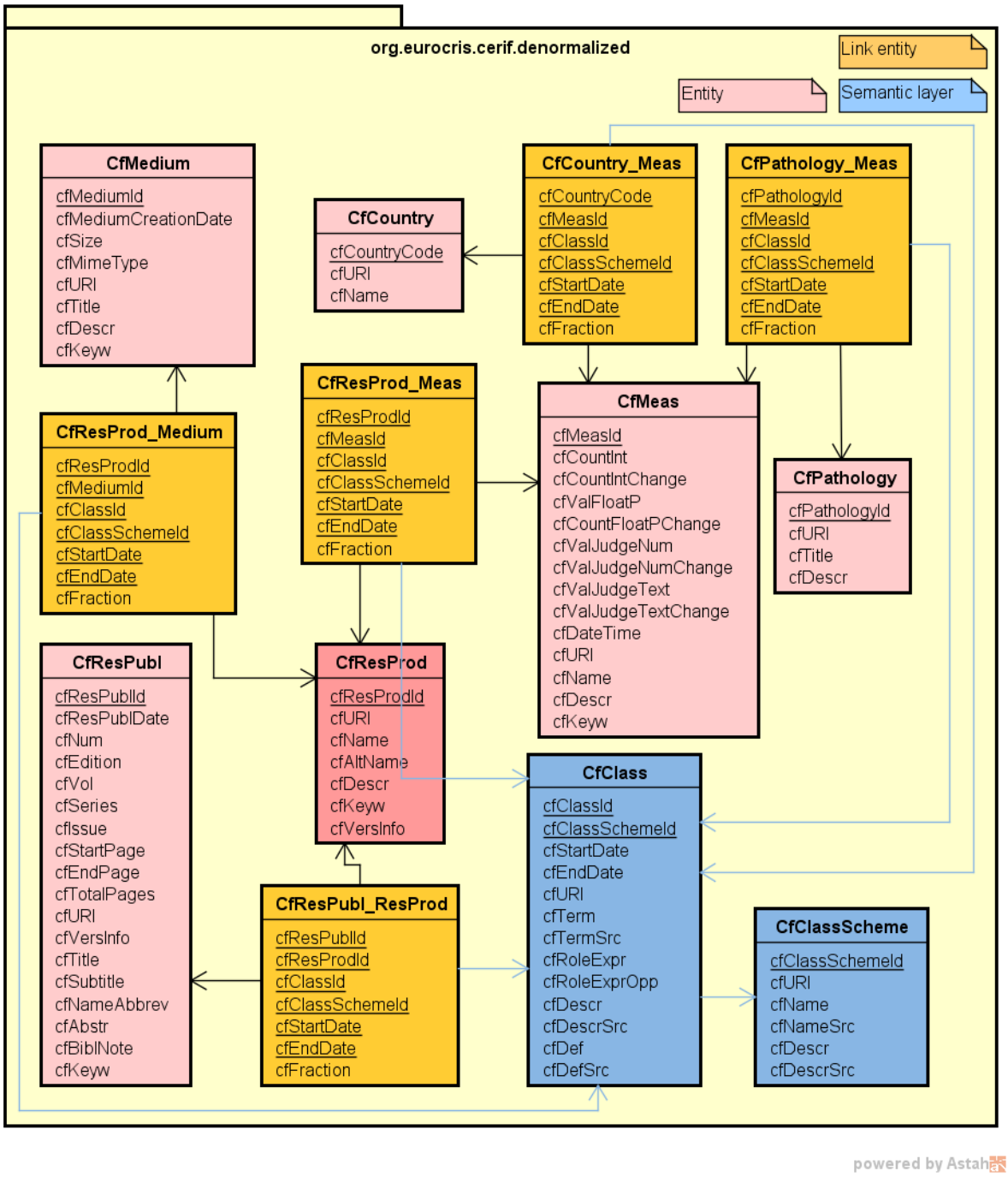
CERIF entities used for patent

The main entity to represent patents in CERIF is cfResPat.



CERIF entities used for procedure

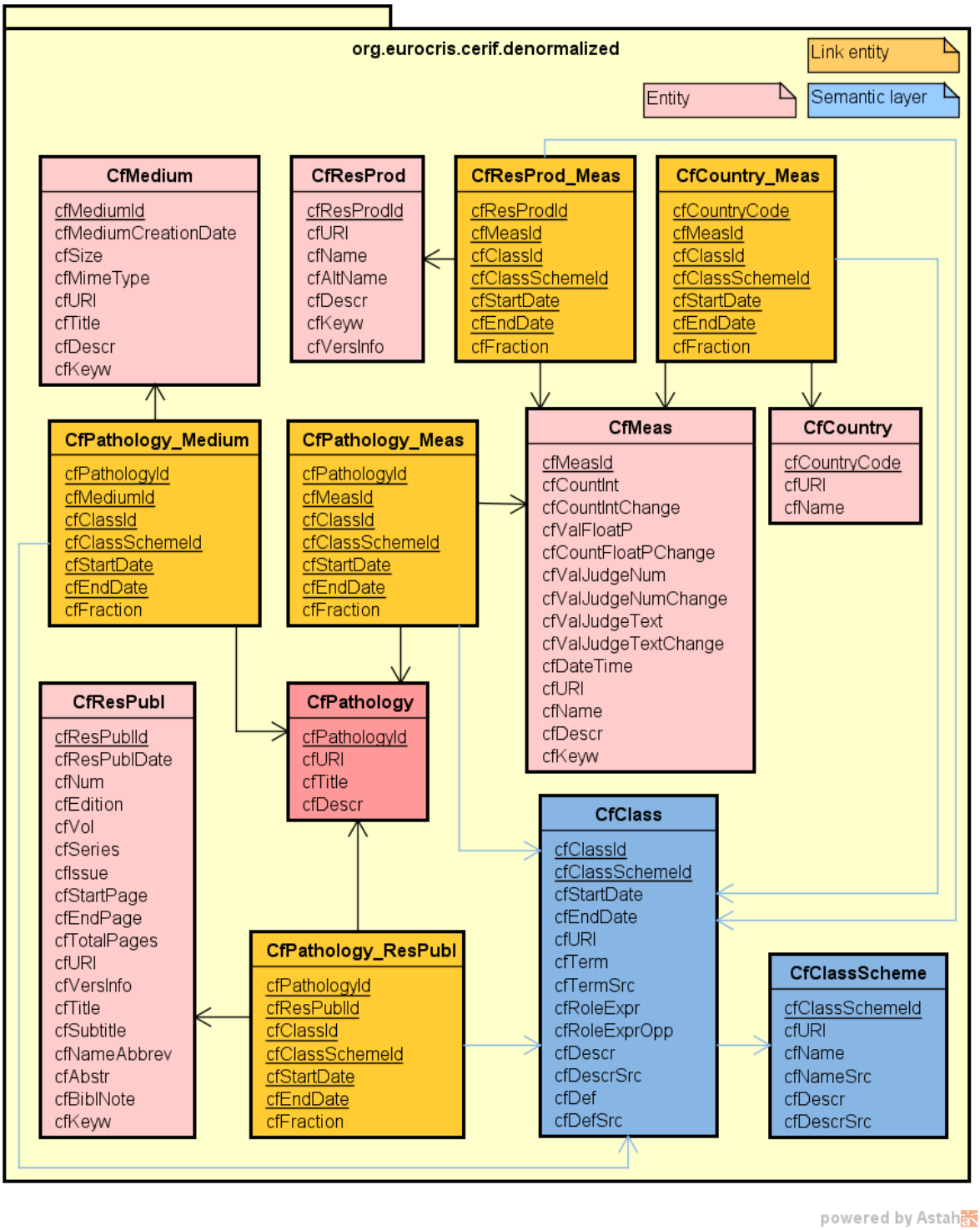
The main entity to represent procedures in CERIF is cfResProd in the context of the knowledge base for IHU.



An additional entity cfCountry_Meas has been added to the original CERIF model to link country to measurements (number of interventions per country for a procedure).

CERIF entities used for pathology

The CERIF extension entity cfPathology is used as the main entity to represent pathology.



The same additional entity cfCountry_Meas has been used (number of cases per country for a pathology).

The CERIF extension is used to manage pathologies and all links to other CERIF entities.