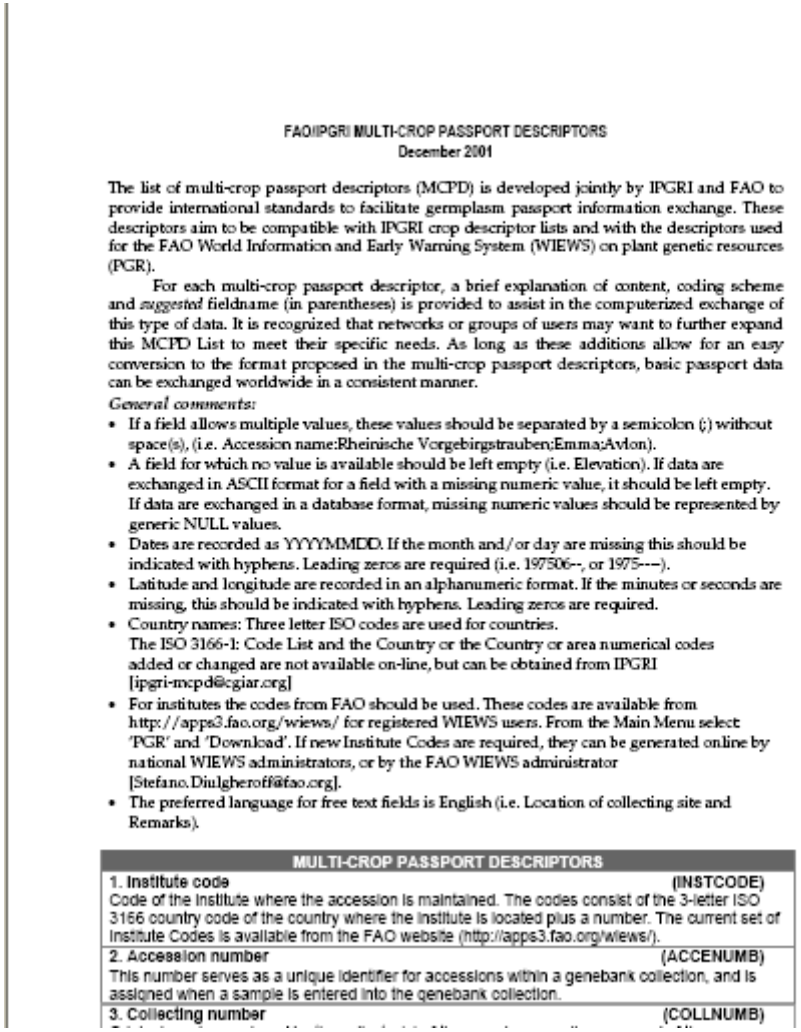
GCP Passport Data Quality Improvement & Assessment Workshop - 3-5 July 2007, Maccarese

Computerization of the MCPD

Tom Hazekamp
Bioversity International

MCPD

- ❑ Flat-file approach
- ❑ Simple format
- ❑ Facilitate adoption



FAO/IPGRI MULTI-CROP PASSPORT DESCRIPTORS
December 2001

The list of multi-crop passport descriptors (MCPD) is developed jointly by IPGRI and FAO to provide international standards to facilitate germplasm passport information exchange. These descriptors aim to be compatible with IPGRI crop descriptor lists and with the descriptors used for the FAO World Information and Early Warning System (WIEWS) on plant genetic resources (PGR).

For each multi-crop passport descriptor, a brief explanation of content, coding scheme and *suggested* fieldname (in parentheses) is provided to assist in the computerized exchange of this type of data. It is recognized that networks or groups of users may want to further expand this MCPD List to meet their specific needs. As long as these additions allow for an easy conversion to the format proposed in the multi-crop passport descriptors, basic passport data can be exchanged worldwide in a consistent manner.

General comments:

- If a field allows multiple values, these values should be separated by a semicolon (;) without space(s), (i.e. Accession name:Rheinische Vorgebirgstrauben;Emma;Avlon).
- A field for which no value is available should be left empty (i.e. Elevation). If data are exchanged in ASCII format for a field with a missing numeric value, it should be left empty. If data are exchanged in a database format, missing numeric values should be represented by generic NULL values.
- Dates are recorded as YYYYMMDD. If the month and/or day are missing this should be indicated with hyphens. Leading zeros are required (i.e. 197506--, or 1975---).
- Latitude and longitude are recorded in an alphanumeric format. If the minutes or seconds are missing, this should be indicated with hyphens. Leading zeros are required.
- Country names: Three letter ISO codes are used for countries.
  The ISO 3166-1: Code List and the Country or the Country or area numerical codes added or changed are not available on-line, but can be obtained from IPGRI [ipgri-mcpd@cgiar.org]
- For institutes the codes from FAO should be used. These codes are available from http://apps3.fao.org/wiews/ for registered WIEWS users. From the Main Menu select 'PGR' and 'Download'. If new Institute Codes are required, they can be generated online by national WIEWS administrators, or by the FAO WIEWS administrator [Stefano.Diulgheroff@fao.org].
- The preferred language for free text fields is English (i.e. Location of collecting site and Remarks).

MULTI-CROP PASSPORT DESCRIPTORS	
1. Institute code Code of the institute where the accession is maintained. The codes consist of the 3-letter ISO 3166 country code of the country where the institute is located plus a number. The current set of Institute Codes is available from the FAO website (http://apps3.fao.org/wiews/).	(INSTCODE)
2. Accession number This number serves as a unique identifier for accessions within a genebank collection, and is assigned when a sample is entered into the genebank collection.	(ACCENUMB)
3. Collecting number	(COLLNUMB)

MCPD (use)

- ECPGR
  - Crop networks
- EURISCO
  - MCPD + 6

EURISCO DESCRIPTORS	
0. National Inventory code	**(NICODE)**
Code identifying the National Inventory; the code of the country preparing the National Inventory. Exceptions are possible, if agreed with EURISCO such as NGB.	
Example: `NLD`	
1. Institute code	**(INSTCODE)**
FAO Institute Code of the institute where the accession is maintained.	
Example: `NLD037`	
2. Accession number	**(ACCENUMB)**
This number serves as a unique identifier for accessions within a genebank collection, and is assigned when a sample is entered into the genebank collection.	
Example: `CGN00254`	
3. Collecting number	**(COLLNUMB)**
Original number assigned by the collector(s) of the sample, normally composed of the name or initials of the collector(s) followed by a number. This number is essential for identifying duplicates held in different collections.	
Example: `FA90-110`	
4. Collecting institute code	**(COLLCODE)**
Code of the Institute collecting the sample. If the holding institute has collected the material, the collecting institute code (COLLCODE) should be the same as the holding institute code (INSTCODE).	
Example: `NLD037`	

MCPD (integration)

- ABCD
  - Mapping of MCPD concepts to ABCD 2.06

MCPD (use)

Generation Challenge Programme (GCP)

	Domain modeling	Data Templates
2005	Initial model explicit	Passport data template based on MCPD + >50 other concepts

MCPD (use)

Generation Challenge Programme (GCP)

Section	Description	Notes
0) Meta data provider	Basic information on the data provider	Mandatory
1) General Passport Data	Basic set of common passport descriptors (adaptation of MCPD[1]) and local uses, special plant characteristics and stress of specific interest to GCP.	Some data should always be submitted.
2) Collecting Location data	Data related to the collecting site and soil characteristics.	Some of these data to be submitted for collected germplasm.
3) Additional Passport Data	Any additional descriptors defined by the data provider that could be of particular interest for the GCP. These could include additional soil or climate characteristics etc.	Optional.
4) Institute codes and decoded values	List of institute codes used in passport data sections and their corresponding decoded name and addresses.	Some data should always be submitted.

MCPD (use)

Generation Challenge Programme (GCP)

2005 Passport Data Template:

- ❑ For the most part still flat file format

MCPD (use)

- Fairly wide adoption
- Flat file format easy

BUT

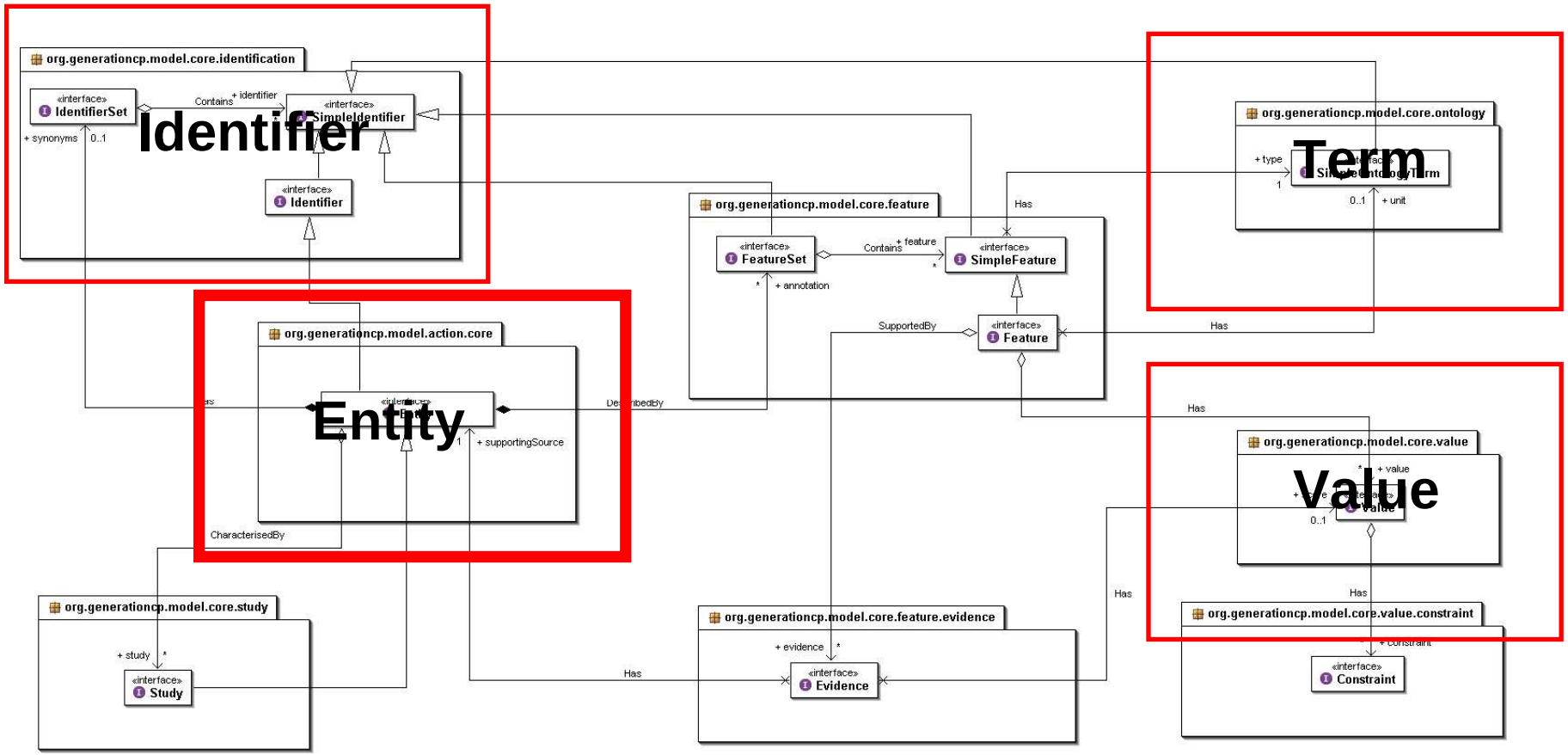
- Potential problems with repeating values
- Many customized schema's incorporating MCPD in use (SINGER, GCP,GPG2)

MCPD (use)

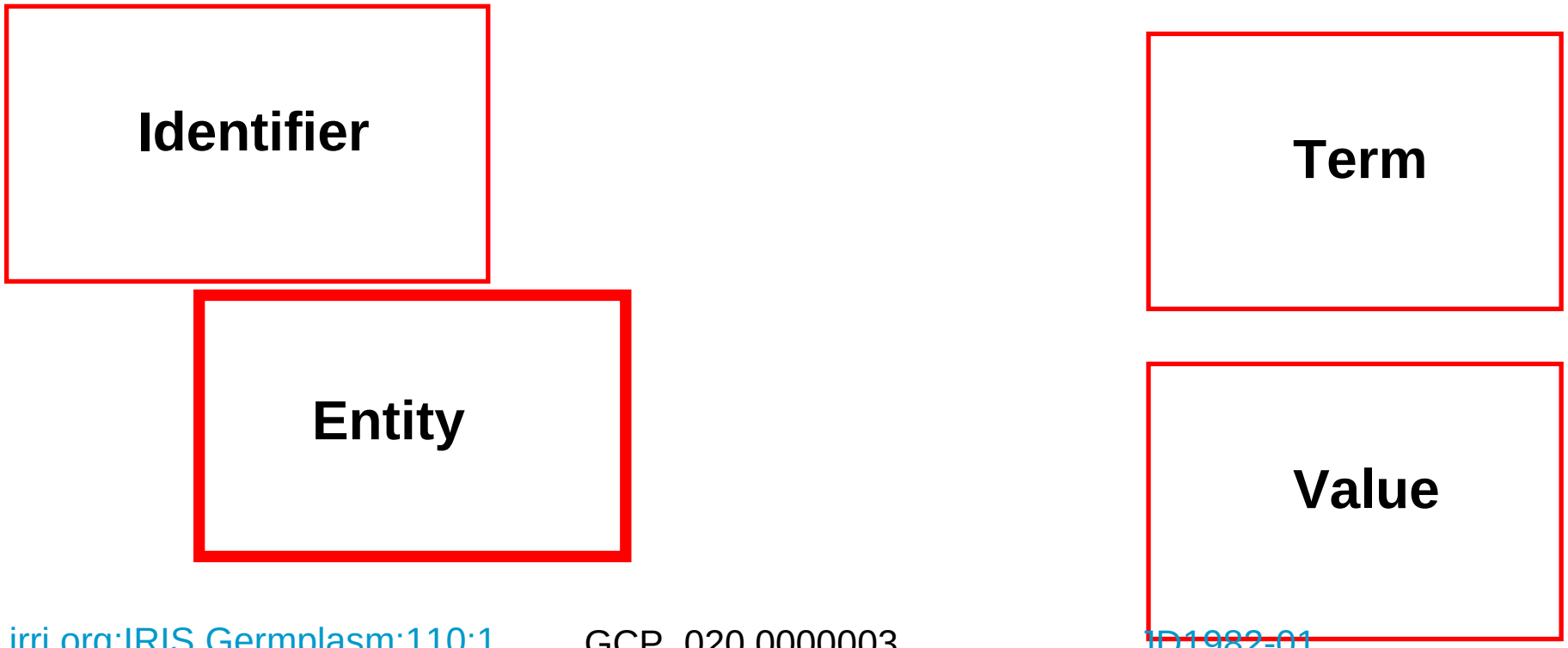
Generation Challenge Programme (GCP)

	Domain modeling	Data Templates
2005	Initial model explicit ‘Hardcoded’	Passport data template based on MCPD + >50 other concepts
2007	Parameterized model + ontologies	

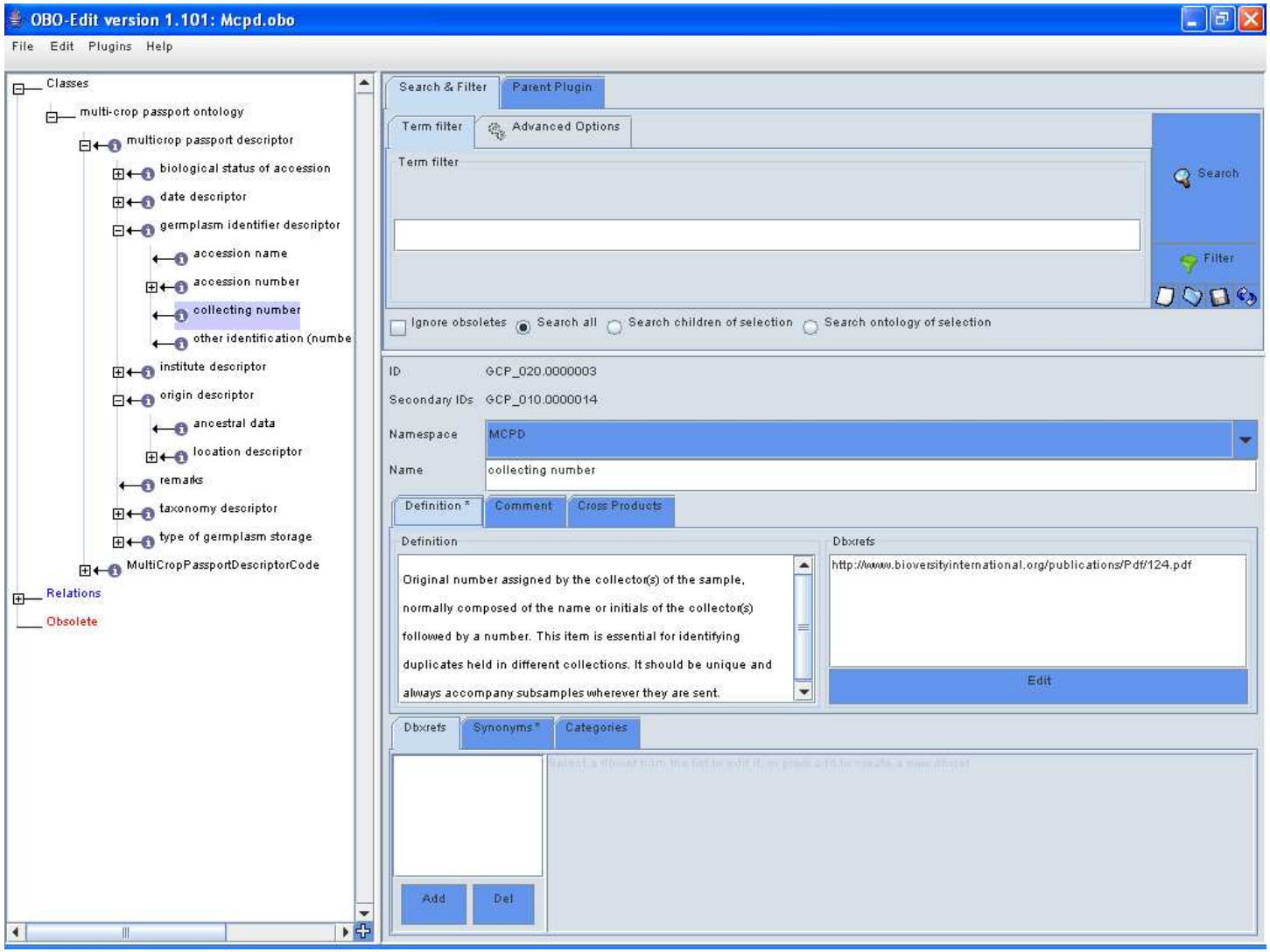
Generation Challenge Programme (GCP)



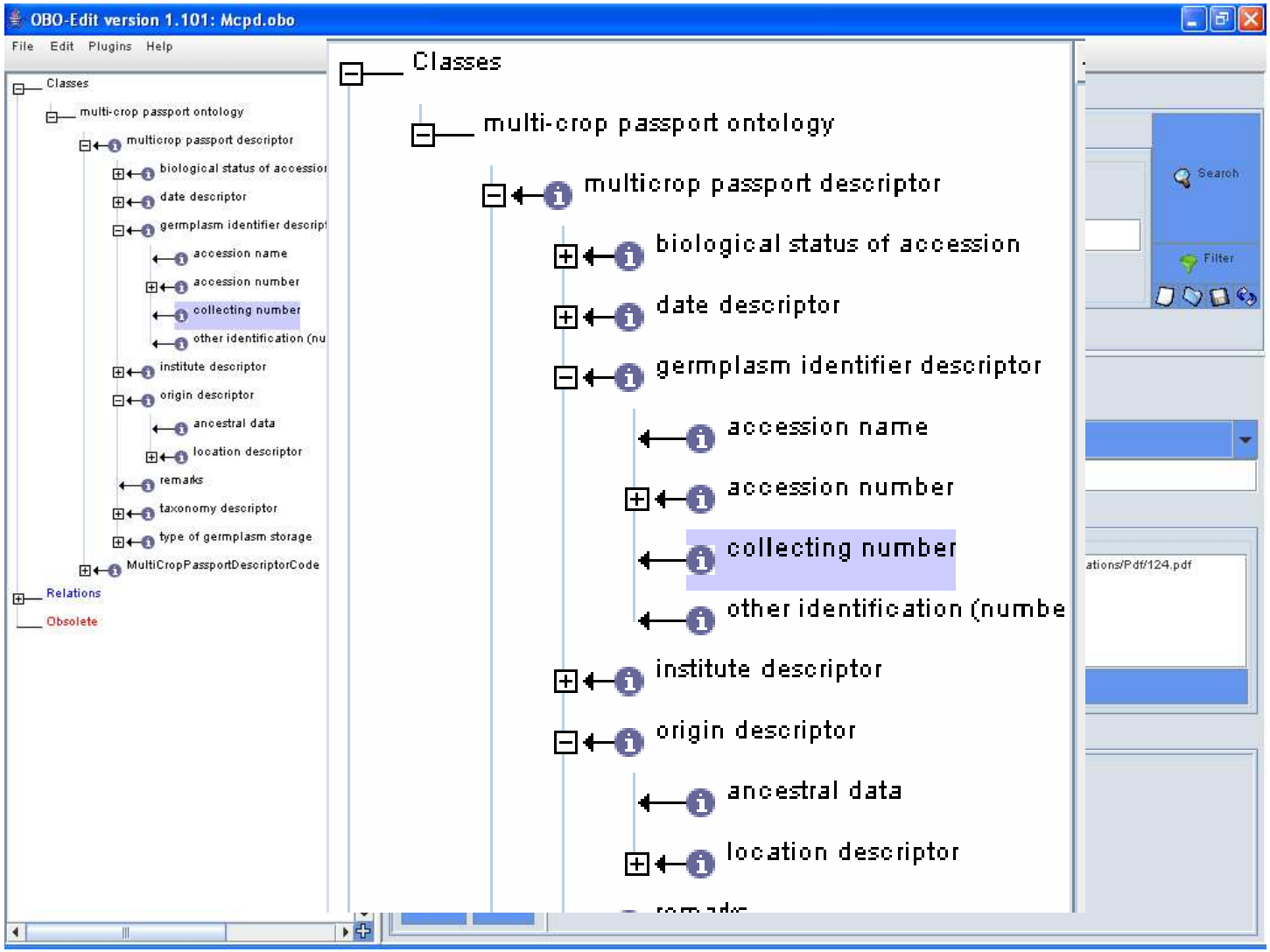
Generation Challenge Programme (GCP)



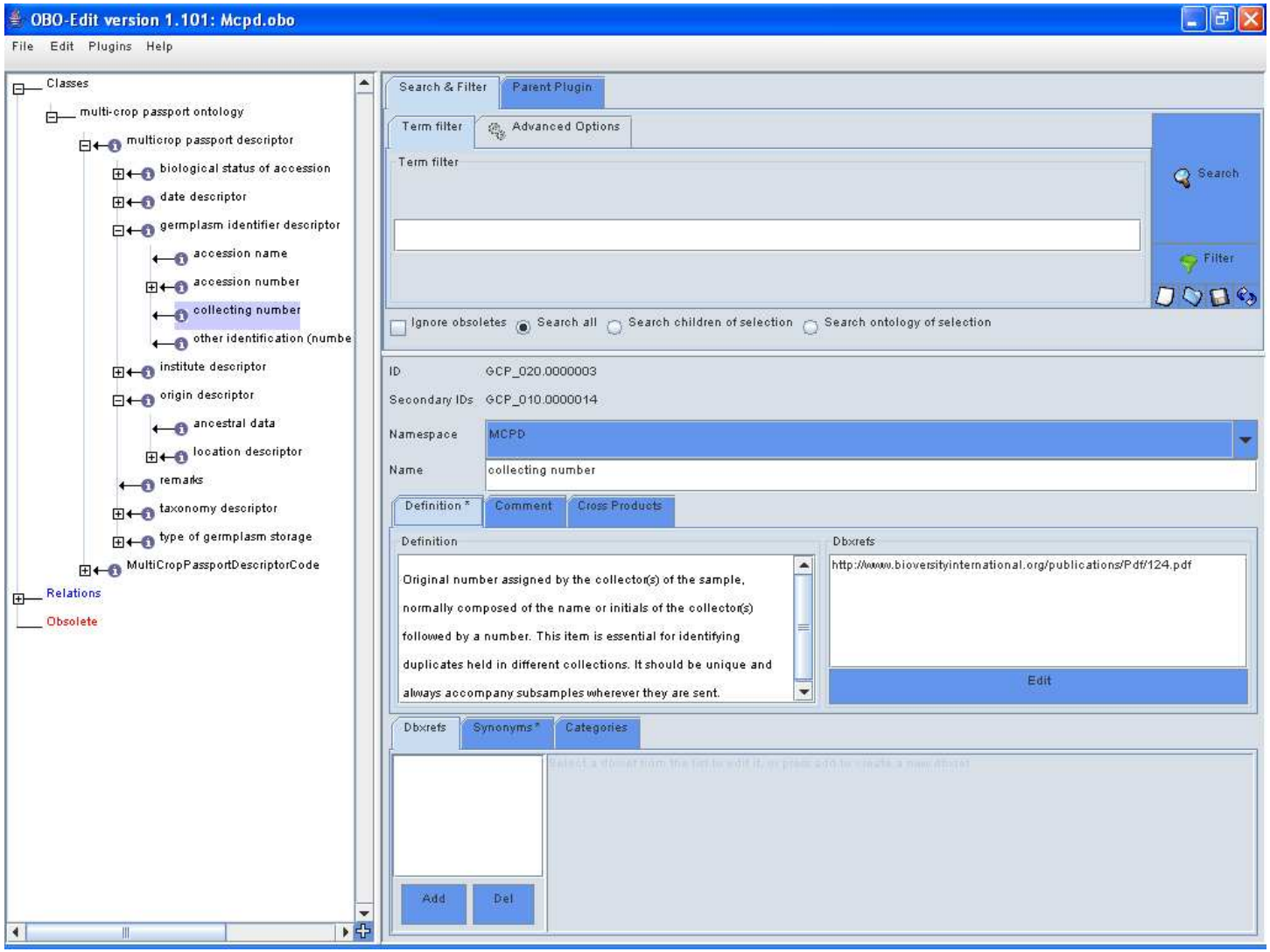
MCPD Ontology



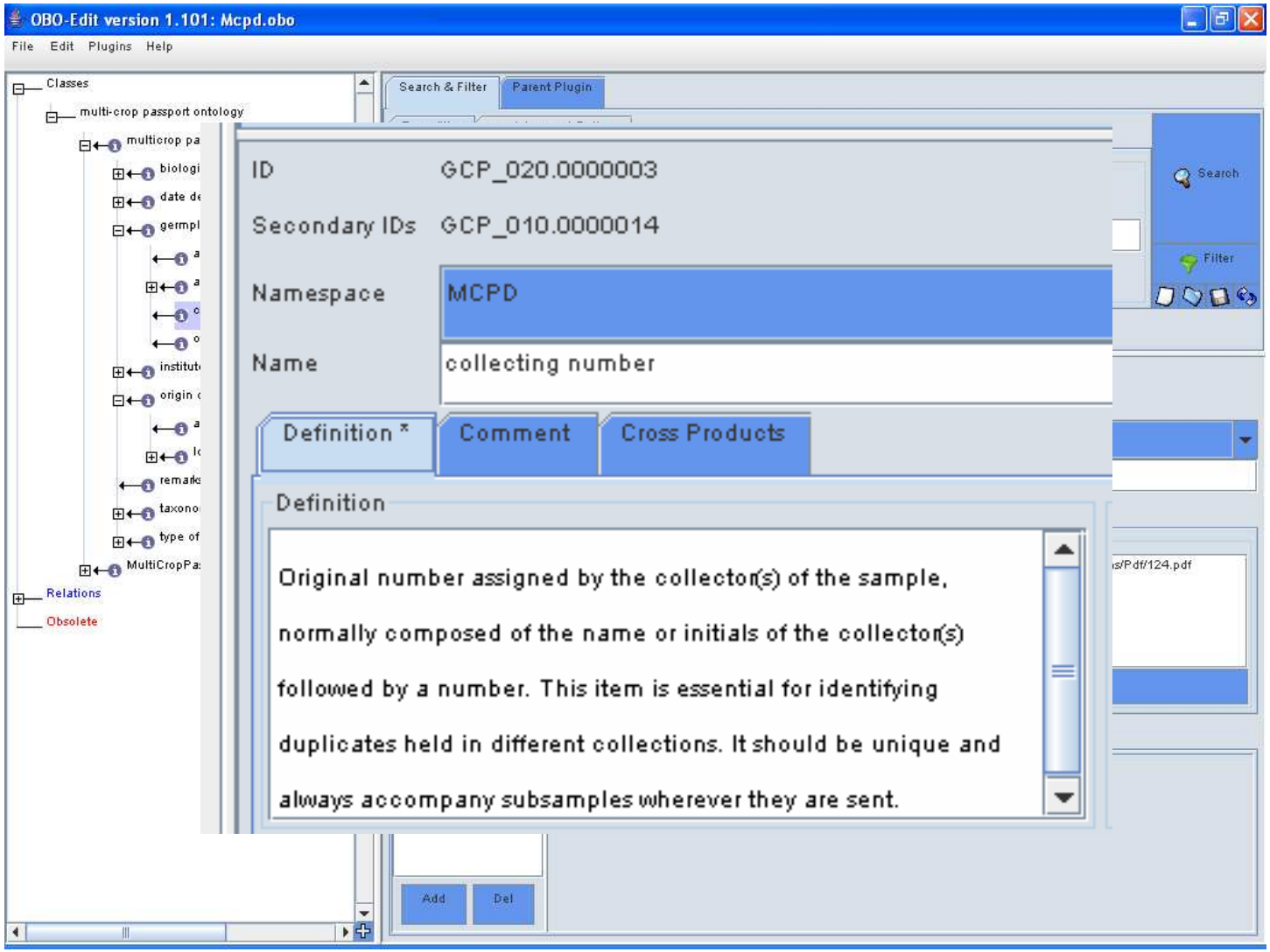
MCPD Ontology



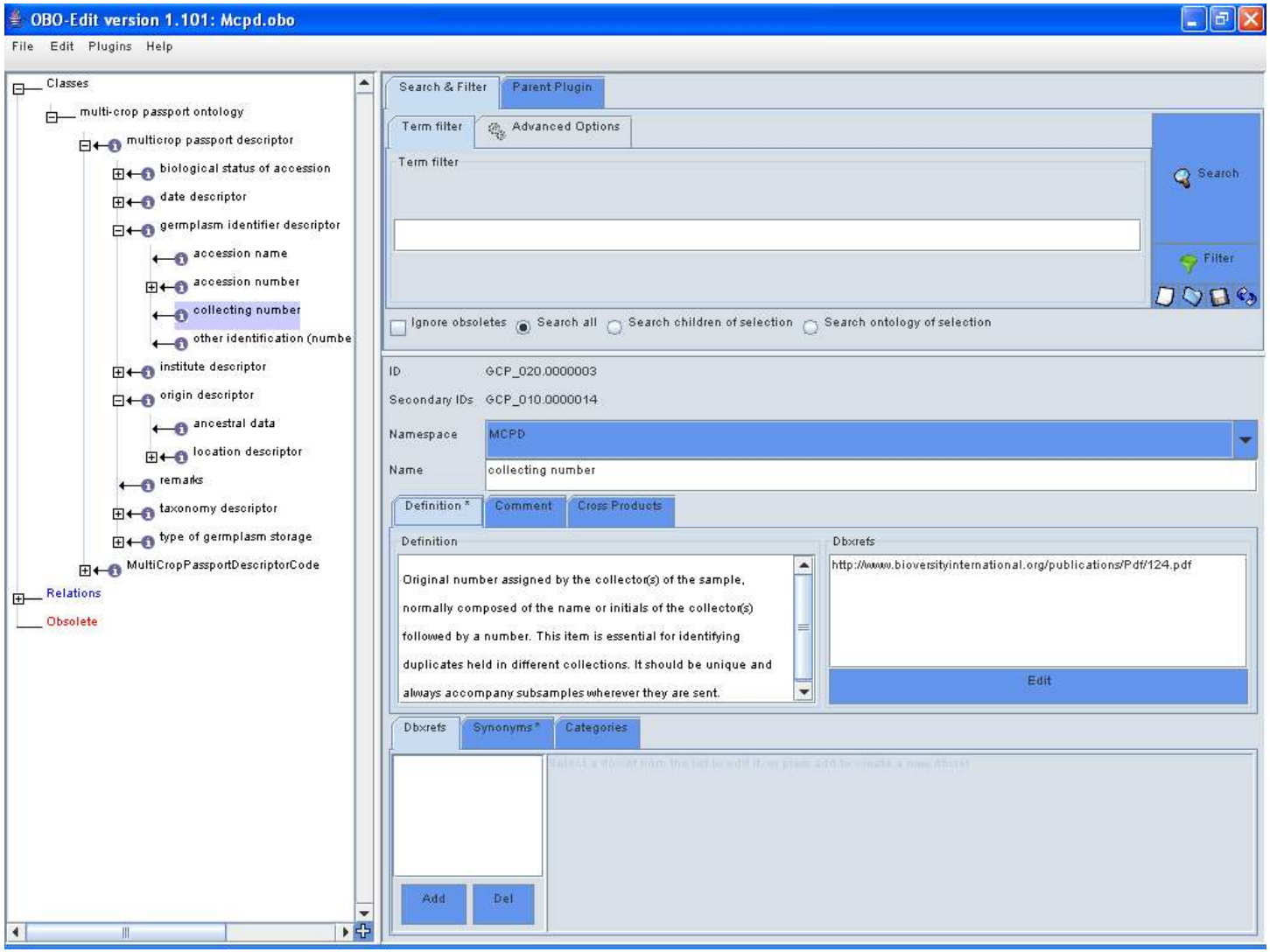
MCPD Ontology



MCPD Ontology



MCPD Ontology



MCPD revision

- ❑ Single exchange format for all passport data
- ❑ Consistent & extendable
- ❑ Data items are atomic (1 datum, 1 row)
- ❑ Overhead, but no empty cells
- ❑ Conversion to format is consistent
- ❑ Customized assembly from format is facilitated
- ❑ Data quality control facilitated

- ❑ Development of coherent, controlled vocabularies
- ❑ Ontology mapping to other domains

MCPD (adoption)

Generation Challenge Programme (GCP)

	Domain modeling	Data Templates
2005	Initial model explicit	Passport data template based on MCPD + >50 other concepts
2007	Parameterized model + ontologies	Revision Passport Templ. Sept-Oct 2007

Concluding remarks

- MCPD well established
- Requires regular revision to stay up-to-date
- Content
  - New concepts
  - Data quality issues
- Efficiency
  - Flat file format restrictive-> flexibility needed
  - Many customized schema’s inc. MCPD ->generic