The ISDC concept for long-term sustainability of geoscience data and information

Bernd Ritschel [(1)], Christian Bruhns [(2)], Ronny Kopischke [(3)], Vivien Mende [(4)], Hartmut Palm [(5)], Sebastian Freiberg [(6)], Lutz Gericke [(7)]

GeoForschungsZentrum Potsdam (GFZ)
Data Center
Potsdam, D-14473 Potsdam, Germany
EMail: rit@gfz-potsdam.de [(1)], cbruhns@gfz-potsdam.de [(2)], roko@gfz-potsdam.de [(3)], vmende@gfz-potsdam.de [(4)], palm@gfz-potsdam.de [(5)], sebast@gfz-potsdam.de [(6)], lg@gfz-potsdam.de [(7)]

ABSTRACT

The GFZ Information System and Data Center (ISDC) for geoscience data is managing data and information from satellite missions, like CHAMP, GRACE and TerraSAR-X as well as from global distributed super conducting gravimeter or GPS ground station networks. Not only the accumulated amount of data is enormous, but also the daily increase rate and the variability of the geoscience data covering geodesy, geophysics and atmospheric research. After the launch of the new ISDC portal in March 2006 more than 250 different product types can be managed using one system only. The integrated ISDC user management is controlling the data access for approximately 1500 registered national and international user groups. More than 15 Million data sets at a subsumed volume of 10 TByte are long-term archived using a tree-tier storage architecture which is providing both a long-term preservation of data and an online access to the data files. The management of the data is based on product type and product related metadata. Parent and child Directory Interchange Format (DIF) XML metadata are transformed and stored in relational database structures. Specific ISDC applications using this information are controlling the ISDC input as well as the ISDC output data flow and the transfer of the validated data files to the different storage media.

Keywords: ISDC, Metadata, CHAMP, GRACE, GCMD, DIF, ISO, Catalog Web Service, OAIS

INTRODUCTION

This paper is dealing with the data life cycle management at the GFZ Potsdam focusing on the situation at the Information System and Data Center (ISDC) portal system, managed by a very small group at the GFZ Data Center. The purpose of this publication is the description of important results concerning the analysis of the development of the ISDC system and the existing status of the ISDC portal always in respect to the responsibility for long-term sustainability of geoscience data and information. Major experiences based on the every day ISDC business as well as conclusions for further conceptual and technical system developments and the organization of operational data management services are completing the paper. The development of the idea for a professional data and information management in form of the ISDC concept was driven by the geodetic community at the GFZ, beginning with the first works for the realization of the CHAMP satellite project in the mid 1990th.

DEVELOPMENT OF THE CHAMP ISDC

Already within the design phase (B) and especially during the implementation phase (C/D) of the GFZ CHAMP satellite project, the need for a professional data management, which should cover almost the complete life time of data from the generation of measured values by sensors, transmitting and receiving

of digital data to the processing, storage and distribution of data files, had been one main objective of the project. The importance of this fact is reflecting figure 1, where the Science Data System and the Information System and Data Center tasks are represented in own work package structures.

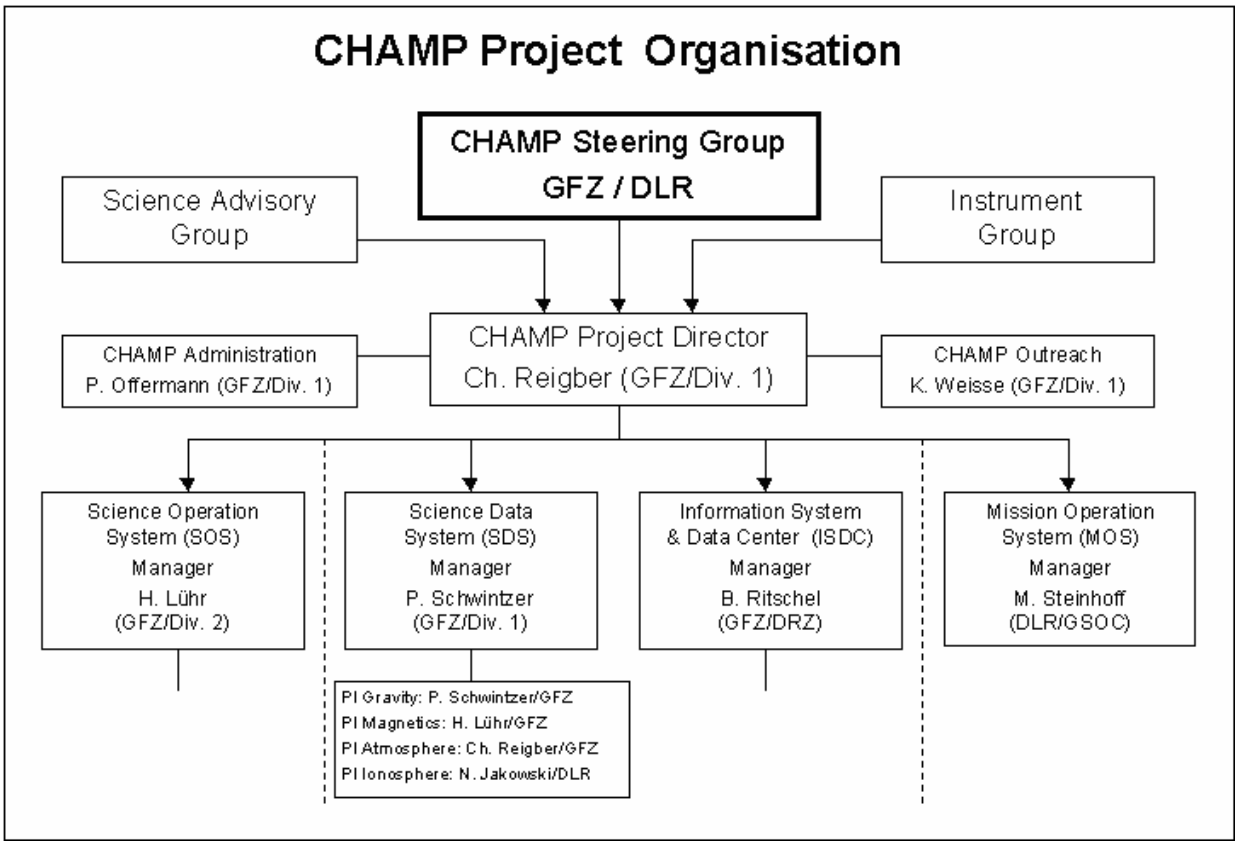

Figure 1: CHAMP Project Organisation

From the beginning of the CHAMP project it was planed, not only to develop just an archive system for the CHAMP data with anonymous FTP access to validated, calibrated and processed data files, called products (figure 3) but also to establish a corresponding clearinghouse containing a product type dependent retrievable catalog system. The name for the CHAMP data management and catalog system: CHAMP Information System and Data Center (CHAMP ISDC) was born [1]. Following to the style of a complex three-tier software system architecture model, figure 2 shows, the main components of the ISDC as well as the basic data and information flows to and from the ISDC system. The Operational System (OS) component which is behind the three layer structure, consisting of the Graphical User Interface (GUI), an application layer and the Product Archive System (PAS) layer, is networking the different parts and is responsible for the different ISDC product management processes done by data pump modules. The product transfer from the data provider to the ISDC system as well as the product transfer to the users is realized by appropriate authorized ISDC FTP services. The GFZ Computing Center's long-term Hierarchical Storage Management (HSM) system and the ISDC-own Online Product Archive (OPA) compose the ISDC PAS layer. Above this part such application layer components like clearinghouse, data warehouse and Product Ordering System (POS) are located (figure 2). Whereas the clearinghouse contains the complete ISDC product catalog realized by product dependent metadata, the data warehouse component is integrating product type dependent metadata as well as other sources of information (see chapter Conclusion and Outlook). Product type dependent as well as product dependent metadata are stored in tables and handled by a relational data base management system. The POS component is necessary because of two important system requirements and constraints respectively. On one hand the long-term HSM as well as the online OPA data storage systems are located behind the firewall within the intranet, on the other hand only registered and authorized users are allowed to

access to the data. Using the ISDC system, there is no anonymous download of data possible. In order to meet these requirements a sophisticated user management module and user right depending data access controlling mechanism had to be developed too. The upper layer of the ISDC architecture is the GUI layer. Most of the interaction between users and the ISDC system is realized by dynamically generated web browser pages containing textual and graphical information. New users, who are really interested in downloading data files, are able to register at the ISDC.

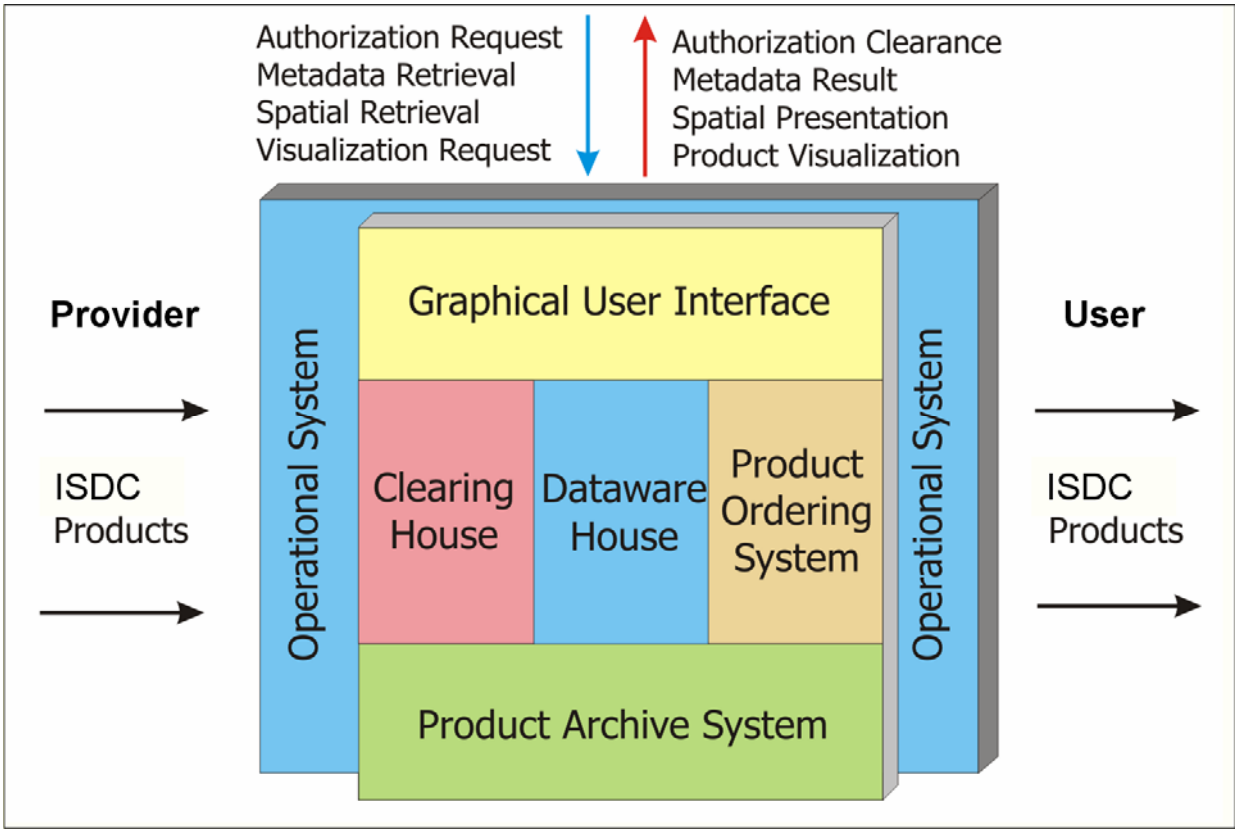

Figure 2: Three-tier CHAMP ISDC system architecture model

The user name is freely selectable, whereas the password is created by the system, but can be changed by the user later on during the extended registration process at the new ISDC portal. According to the CHAMP data policy, all data must be used for research reasons only. In order to meet this requirement, users have to provide application information. This information is used by the project administration in order to take a decision about data access grants related to different product types and products. Registered and authorized users can use various methods in order to search for required data using the ISDC product catalog. Product dependent tailored retrieval forms, qualified by search attributes provide return pages containing the query results added by features in order to start the product ordering process, after activating the search process. There is also a batch request for bulk data implemented, which is realized by product request list files. The new ISDC portal even is providing a product browser tool with "shopping cart" functionality where the user is searching for data within virtual directory trees. For all products, whose metadata are containing qualified spatial information other than a global spatial reference, a map server based spatial search tool is usable too.

ISDC PRODUCT PHILOSOPHY

The complete management and handling of all ISDC data is based on product type and product related metadata. After an intensive search and validation process for an appropriate metadata structure in order to describe and manage the data, NASA's Directory Interchange Format standard was chosen. Whereas

this metadata standard is excellent for describing product types (data directories, data sets, data series), it was necessary to extend the DIF standard in order to describe specific features of single products (individual data files, granular files, granules), using data file dependent metadata too. Figure 3 illustrates the ISDC product philosophy, which is valid for all ISDC products. All relevant information about the ISDC products, often only stored in file headers or control and configuration files or hidden in processing software or sometimes only written on a sheet of paper or available in the memory of scientist' brains, like unique identifier and title, different kind of science and technical keywords, summarized description and references, information about data quality, temporal and spatial coverage as well as information about sensors and instruments but also file size and md5 sum are structured preserved in the extended ISDC DIF metadata files.

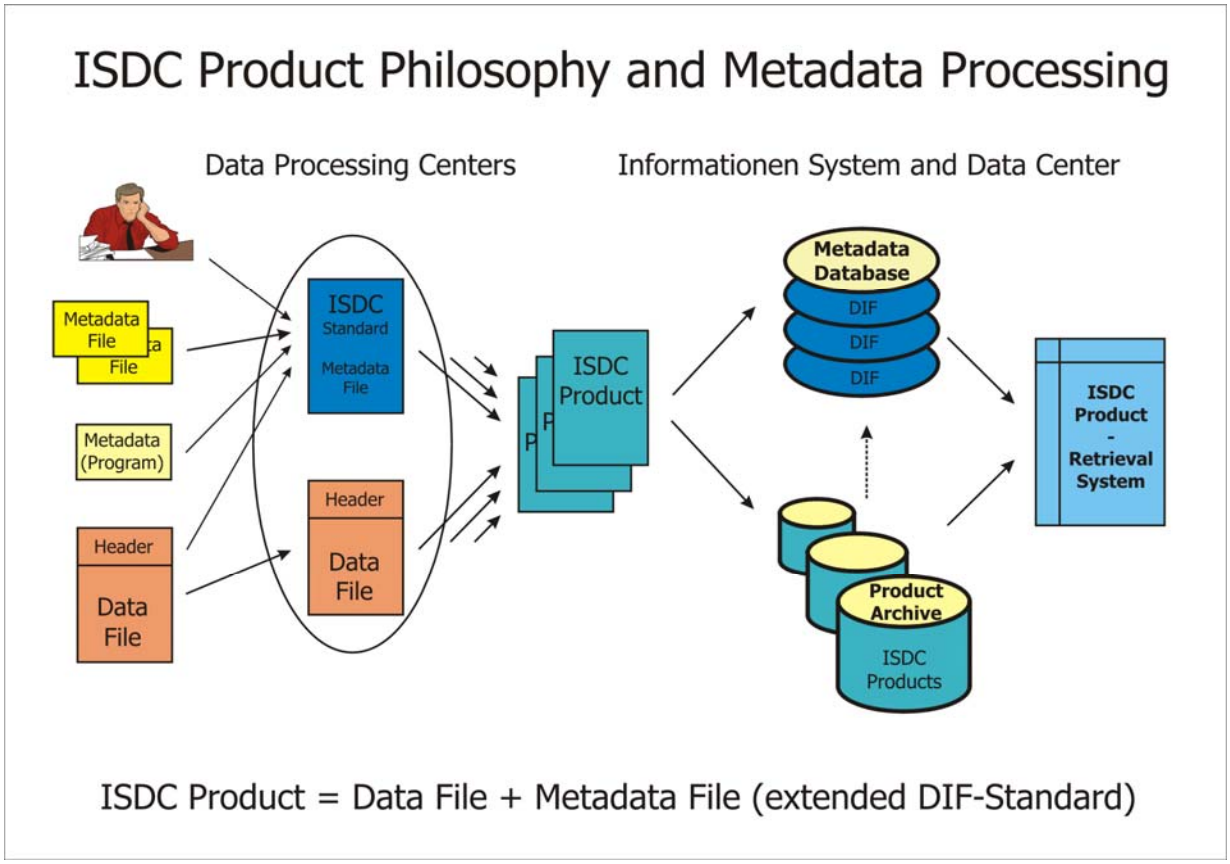

Figure 3: ISDC product philosophy

Following this ISDC product philosophy, meanwhile almost 300 different product types can be managed using standardized software only. A non-stop ISDC product check-in process is not only continuously parsing every new metadata file but also extracting and transferring product type as well as product related meta information into appropriate relational database structures. In order to be sure new products within the ISDC FPT input directory were transferred completely, such data file meta information like file size and md5 sum are checked and compared with appropriate values extracted from the appropriate metadata files. Assuming that the check was successful and valid, both, the data file and the metadata file are transferred to the different parts and directories of the ISDC product archive and appropriate processing flags are set in special control and monitor tables, which are responsible for the handling of the product input process. The fact of keeping equal product type meta information in every metadata file is minimized by extending the ISDC product philosophy. The quasi-static product type only metadata are removed from the product metadata files and stored in separate product type related metadata files only once. The product (data file) dependent meta information are stored in appropriate metadata files. The connection between the product type metadata file and appropriate product metadata

files are assured using object-oriented parent-child heritage relations between these different metadata file types (figure 4). According to the enhanced ISDC product philosophy, for each different product type all product type only related metadata are documented in a separate parent DIF metadata file whereas the product (data file) only metadata supplemented by mandatory metadata are documented in separate child DIF metadata files. The step-by-step realization of this enhanced ISDC product philosophy is associated with a change of the DIF metadata format from ASCII text to XML structures.

Figure 4: Enhanced ISDC metadata standard

The constraints for a transformation of ASCII text based DIF content into XML based DIF structures are relatively low because of similar structures of both metadata format types. The same hierarchical order of metadata elements is used furthermore. There is additional effort necessary in order to define the different product type dependent XML schemata. In order to handle XML documents, changes within the data pump software must be realized. The enhanced ISDC product philosophy has been used already for the new GFZ cooperation projects GPS-PDR (GPS reprocessing) and TerraSAR-X. Examples for a parent and a child DIF XML document (shortened) are shown in figure 5 and figure 6. The parent DIF XML metadata document (figure 5) is describing the CHAMP (orbit/gravity, processing level 3) Rapid Science Orbit product (CH-OG-3-RSO), whereas the metadata of an appropriate product, like e.g. CH-OG-3-RSO+CTS-CHA_2000_219_10.dat, is documented in a child DIF XML file (figure 6).

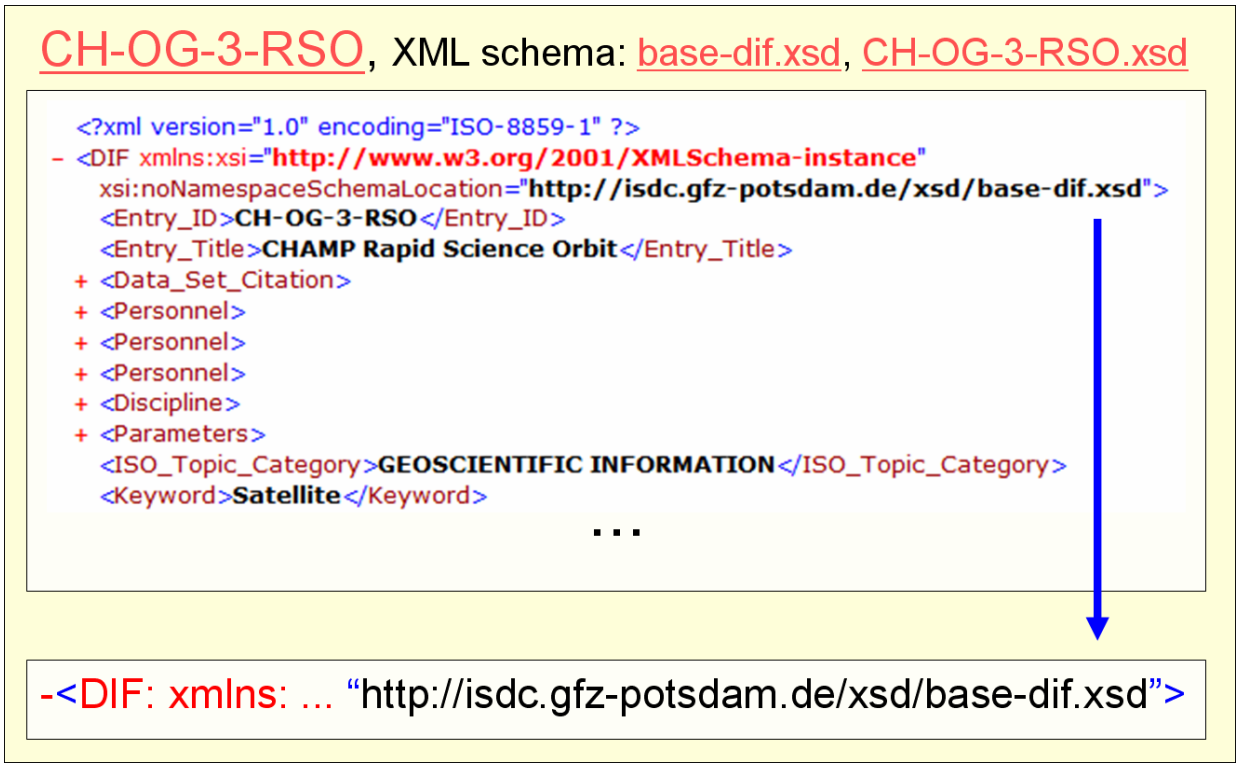

Figure 5: CH-OG-3-RSO parent DIF XML document

All parent DIF XML documents are based on the ISDC base-dif.xsd XML schema definition file, which is containing the document structure and defining all XML elements, attributes and partially valid attribute values too. Such elements like <Entry_ID>, <Entry_Title> or the sequence <Parameters> are

mandatory and must be used in child DIF documents as well. In order to make allowance for the fact of the great variability of product specific metadata, different child DIF XML schemata are used in order to define product type dependent XML elements. For example, all CHAMP, GRACE or TerraSAR-X atmospheric or ionospheric occultation products are described amongst others by such specific metadata attributes, like occultation satellite, reference satellite and appropriate satellite channels.

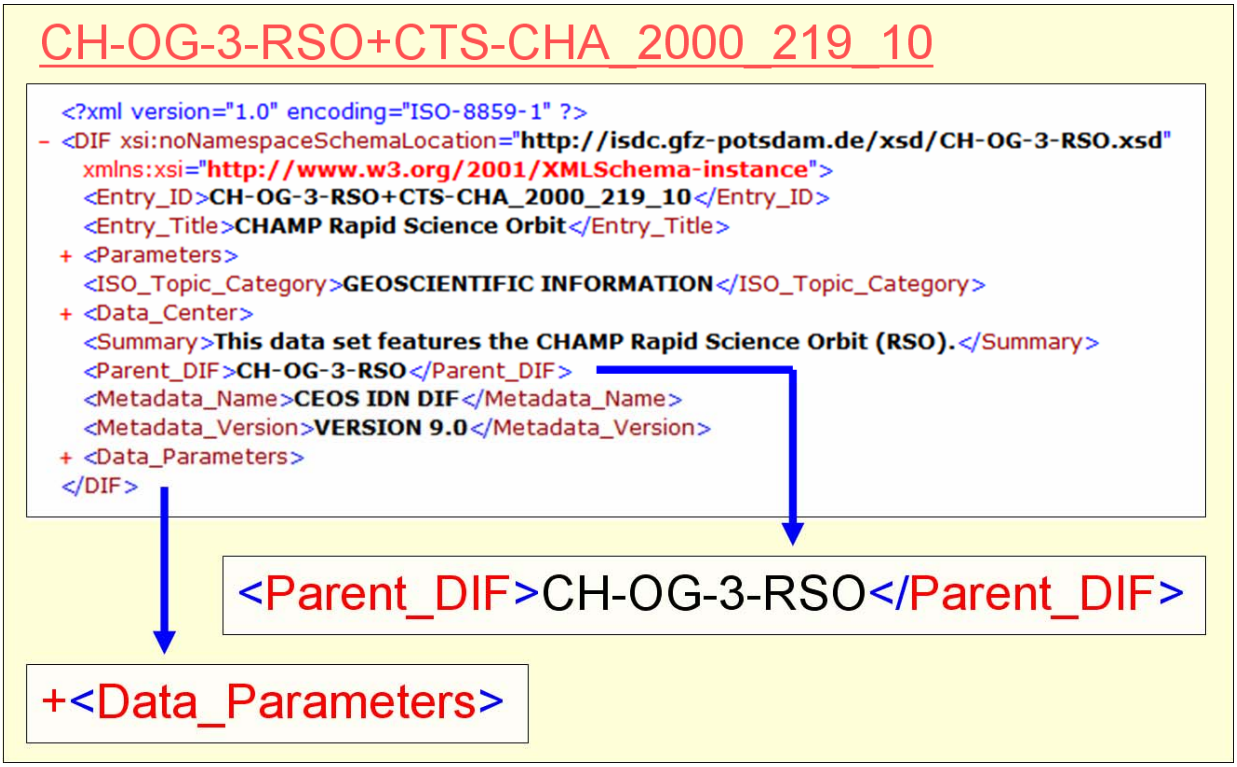

Figure 6: CH-OG-3-RSO+CTS-CHA_2000_219_10 child DIF XML document

Both, the inheritance relation between a parent DIF and an appropriate child DIF and the enhancement of the GCMD DIF standard by the <Data_Parameters> XML sequence are illustrated in figure 6. The inheritance relation is represented by the value of the <Parent_DIF> element. The example shows the relation between the CH-OG-3-RSO+CTS-CHA_2000_219_10 child DIF metadata document and the related CH-OG-3-RSO parent DIF metadata file. The <Data_Parameters> XML sequence is containing elements regarding to the product content and access grants, like e.g. <Data_Entry_ID> (unique ID of the data file), <Start_Time> and <Stop_Time> (temporal coverage of the data), <Reference_Frame> and <Data_Access> as well as elements regarding to technical details of the data file, like e.g. <Name> (file name), <Size>, <MD5> (md5 sum) and <Format>. This state-of-the-art XML-based ISDC product type and product related metadata concept offers a consistent and standardized management of most different data sets or series and appropriate single or granule data files. Even more, the usage of XML for the description of metadata and the close relatedness between the DIF and the ISO 19115 metadata standard [2] open the door for real interoperability using catalog and data web services.

ISDC STORAGE MANAGEMENT

According to ISDC data storage objectives the ISDC data pumps are able to transfer all products to the ISDC OPA and to the GFZ HSM system as already described in chapter “Development of the CHAMP ISDC”. Subject to the condition that enough storage capacity is available at the ISDC OPA in form of hard disc RAID systems and in form of tapes related to the GFZ HSM system, the original product and the first copy of a product are stored at the HSM system for long-term archiving whereas the second copy is stored at the ISDC OPA for online access to the products. Both, the ISDC OPA and the GFZ

HSM system are located within the GFZ Intranet. Following this three-tier storage architecture concept a save, secure and sustainable preservation of data should be possible. The substantial experiences of handling unique data from the year 2000 until now, illustrate both, the necessity of a triple storage of data on different media and the usage of different storage systems (OPA and HSM) administrated by different persons. Unfortunately, insufficient funding of storage media and systems on one hand and the lack of qualified manpower on the other hand are vitiating the best concepts and objectives.

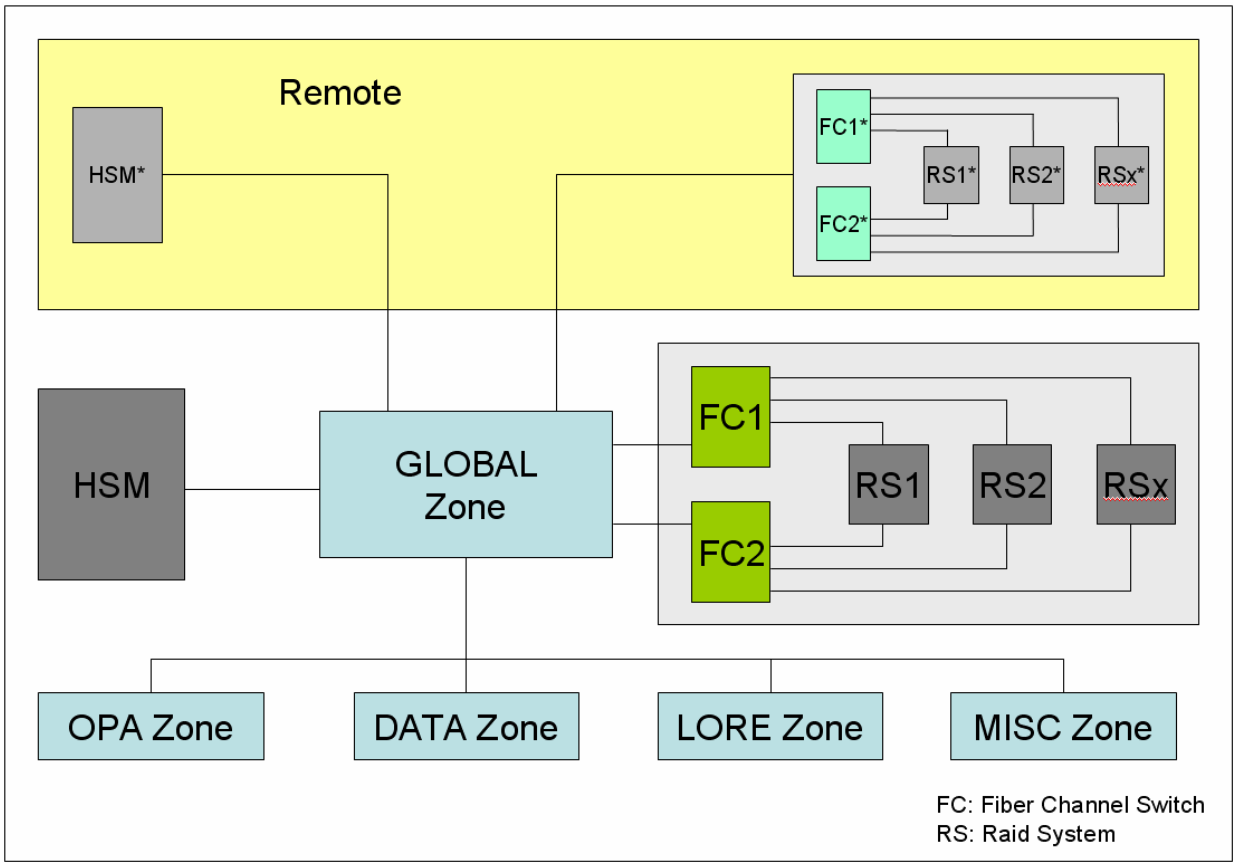

Figure 7: ISDC storage management structure

The new ISDC storage architecture is shown in figure 7. Using the zoning features of the new ZFS file system and the support of logical domains of the SUN Solaris 10 operating system, an easy but powerful storage management system can be created. Derived from the GLOBAL zone, single operating system zones for controlling the OPA, the FTP services DATA and LORE as well as miscellaneous tasks (MISC) can be established. Because the FTP service DATA is used for the connection to the Internet whereas the OPA, LORE and MISC services are used within the Intranet only, different network connections are necessary. The latest version of SUN Solaris is supporting this requirement. In addition to the allocation of different logical domains to the different zones, the full accessibility to the raid system based storage of the OPA, which is connected via redundant fibre channel switches is available for all different zones too. The complete configuration of the ISDC OPA structure consists of a second mirrored fibre channel switches and raid system based storage system on a different location, added by connections to the HSM system. The first part of the new ISDC storage system will be realized still this year. Considering the new ISDC product philosophy combined with the new ISDC storage management structures and techniques, many ideas and recommendations of the Reference Model for an Open Archive Information System (OAIS) are realized already.

CONCLUSIONS AND OUTLOOK

The ISDC system, based on the ISDC product philosophy is supporting the data life cycle management of scientific products related especially to the successful GFZ satellite missions CHAMP and GRACE [3]. Only the development and consequent usage of standardized metadata concepts for the description and handling of data enables the opportunity to manage almost 300 different product types and more than 15 million products. The maintenance of the sustainability of unique data can be guaranteed realizing the new ISDC storage management system. The new ISDC portal, the door to unique data and information is integrating the former separate, project and mission related ISDC systems under one umbrella with uniform access to data and information. The system architecture of the new portal enables the integration of new projects and products as well as the integration of new data retrieval methods using the ISDC catalog and the connection to other information systems, data centers and electronic libraries using interoperable catalog web services. The development and enhancement of methods for the best description of data using user driven Web 2.0 technologies like tagging, social networking and mashup techniques are useful for an inter-domain usage of geoscientfic data available at the GFZ ISDC.

REFERENCES

[1] – Ritschel, B., Braune, S., Behrends, K., Freiberg, S., Kopischke, R., Palm, H., Schmidt, A., Schneider, M. (2003): CHAMP-ISDC - Informationssystem und Datenzentrum für geowissenschaftliche Produkte des CHAMP-Satellitenprojekts. - Zeitschrift für Geologische Wissenschaften, 31,1, 21-30.

[2] – Braune, S., Czegka, W., Klump, J., Palm, H., Ritschel, B., Lochter, F. A. (2003): Anwendungen ISO-19115-konformer Metadaten in in Katalogsystemen aus dem Bereich umwelt- und geowissenschaftlicher Geofachdaten. - Zeitschrift für Geologische Wissenschaften, 31,1, 37-44

[3] – Ritschel, B,. Bendig, A., Palm, H., Flechtner, F., Meyer, U. (2006): Design and operation of the GRACE ISDC - In: Flury, J., Rummel, R., Reigber, Ch., Rothacher, M., Boedecker, G., Schreiber, U. (Eds.), Observation of the Earth System from Space, Springer, 71-82.