The GED4GEM Project: Development of a Global Database for the Global Earthquake Model Initiative

P. Gamba, D. Cavalca
Università degli Studi di Pavia, (Italy)

K. S. Jaiswal
U.S. Geological Survey Golden CO/Synergetics Inc. (USA)

C. Huyck
Imagecat Inc, (USA)

H. Crowley
GEM Foundation (Italy)

In order to quantify earthquake risk of any selected region or a country of the world within the Global Earthquake Model (GEM) framework (www.globalquakemodel.org/), a systematic compilation of building inventory and population exposure is indispensable. Through the consortium of leading institutions and by engaging the domain-experts from multiple countries, the GED4GEM project has been working towards the development of a first comprehensive publicly available Global Exposure Database (GED). This geospatial exposure database will eventually facilitate global earthquake risk and loss estimation through GEM's OpenQuake platform. This paper provides an overview of the GED concepts, aims, datasets, and inference methodology, as well as the current implementation scheme, status and way forward.

Keywords: exposure, global database, data fusion

1. INTRODUCTION

Global datasets, largely extracted from remotely sensed data, do not directly provide the characteristics of buildings and their expected performance under earthquakes. Typically, these data delineate coarse land use classes - such as developed and undeveloped regions - or estimate population based upon land use development patterns observed in moderate resolution imagery. Similarly, the land use information by itself, does not distinguish the structural classes that are necessary for earthquake damage/loss functions. In order to systematically characterize the structural system, it is important to compile the building data related to attributes/features describing building size, shape, configuration, structural material, lateral and vertical force resisting system, etc. The lack of structural information is typically addressed through development of mapping schemes that can make use of existing data such as building census surveys, engineering field surveys, or data inferred using optical remotely sensed imagery.

Building inventory databases at the town, district and/or regional level may exist, however significant efforts are necessary to process and map the raw data before it can be ingested by the Global Exposure Database (GED). Without any guidelines for systematically developing these detailed and aggregated databases or lack of concerted effort to make systemic changes in how the data are compiled (e.g., housing census survey), the prospect of compiling and maintaining a high quality exposure database is bleak. In order to compile global, multi-scale and regional exposure database, significant efforts and resources are required.

State-of-the-art methodologies for development of systematic structural and occupancy-related inventory data and mapping scheme are discussed in ATC-13 (ATC, 1985) and in the NIBS-FEMA funded HAZUS study (NIBS-FEMA, 2009, http://www.fema.gov/plan/prevent/hazus/index.shtm). HAZUS provides a method of categorizing structural occupancy into Model Building Types (MBTs)

for the purpose of loss estimation. Although a few projects have adapted the HAZUS platform for international use (e.g., Australia, Taiwan, Canada), the system is tailored for use in America- primarily because the database architecture is dependent on the US Census and commercial datasets, and the model building type categorization, which are mostly relevant to US construction practices. In order to rapidly estimate the impact of large worldwide earthquakes, the U.S. Geological Survey has recently launched a new product called PAGER (Prompt Assessment of Global Earthquakes for Response, http://earthquake.usgs.gov/pager/). Due to the absence of worldwide building inventory data, the PAGER team developed country-specific statistical distributions of population according to global structural types. Such characterization varies between urban and rural environments and between residential and nonresidential occupancies (Jaiswal and Wald, 2008). The authors compiled data from national censuses, UN datasets, surveys and technical reports taking help from local experts and civil engineers, and developing source-specific mapping schemes. These statistical profiles vary in quality and are dictated by the quality of publicly available structural and inventory information.

The USGS PAGER system addresses exposure at a global scale, providing a critical starting point in that it represents significant progress toward the GED. The PAGER/HAZUS endeavor incorporates best practices critical to the creation of a global dataset for the purpose of assessing risk, including statistical characterization of structural type by country, the leveraging of existing global datasets, and a proven method of extrapolating structural parameters from parcel, or detailed inventory datasets or remotely sensed data.

To improve on these existing state-of-the-art of methodologies, it should also be noted that global datasets are quickly evolving, as demonstrated by the analysis in Gamba and Herold (2009) for global remotely sensed databases related to human settlements, which now include information from moderate resolution optical data but in the near future are likely to be extended to high resolution (Pesaresi *et al.*, 2008) as well as to radar data (Dell'Acqua, 2009). These data bring in important information that no longer can be considered at a single scale, since more and more sensors are providing information at moderate, high and very high *spatial* resolution and moderate to very high *spectral* resolution as well. It is likely that future global datasets will comprise different scales depending on the importance/land cover/focus of the area. In urban areas more details will be obtained and can be incorporated into the GED, while rural areas are likely to be imaged with coarser details and will probably require a different way to integrate data (see for details, Schneiderbauer 2007) for population mapping aimed at computing the social vulnerability component of the risk equation. Datasets like the Global Risk Data Platform (http://preview.grid.unep.ch/) developed by UNEP (United Nations Economic Program) are thus likely to be too coarse in the near future, although they are very important sources of information in the first phase of the development of the GED.

The key attributes of the GED developed by the GED4GEM consortium will thus be as follows:

- **Multiple scales**: the information available and thus the inputs of the GED are for various spatial scales. It is likely that the output of the GED will also be required to be at different scales. Both points call for a system able to store information at multiple geographical scales.
- **Multiple sources**: the GED will take into account and homogenize data coming from census, field surveys, globally available maps, remotely sensed images, and from a variety of other different sources.
- **Multiple inference schemes**: the GED will include multiple ways to infer missing information from available data and proxies, including disaggregation from geographically coarser sources, aggregation from finer ones, combination from multiple datasets.
- **Flexibility with respect to building taxonomy**: the GED will be able to adapt to different (global) taxonomies, regional/local knowledge, additional inputs and information defining building classes according to the local design and construction and architectural practices of a country or a place.

These features will entail 1) reviewing the GIS and remote sensing based datasets to determine where the number of classes can be expanded beyond rural/urban; 2) reviewing all available sources to determine whether assumptions in the development of algorithms for inference mapping can be

expanded, with emphasis on "missing countries"; and 3) applying the developed algorithms to the different datasets collected by the project team to create the final global exposure database.

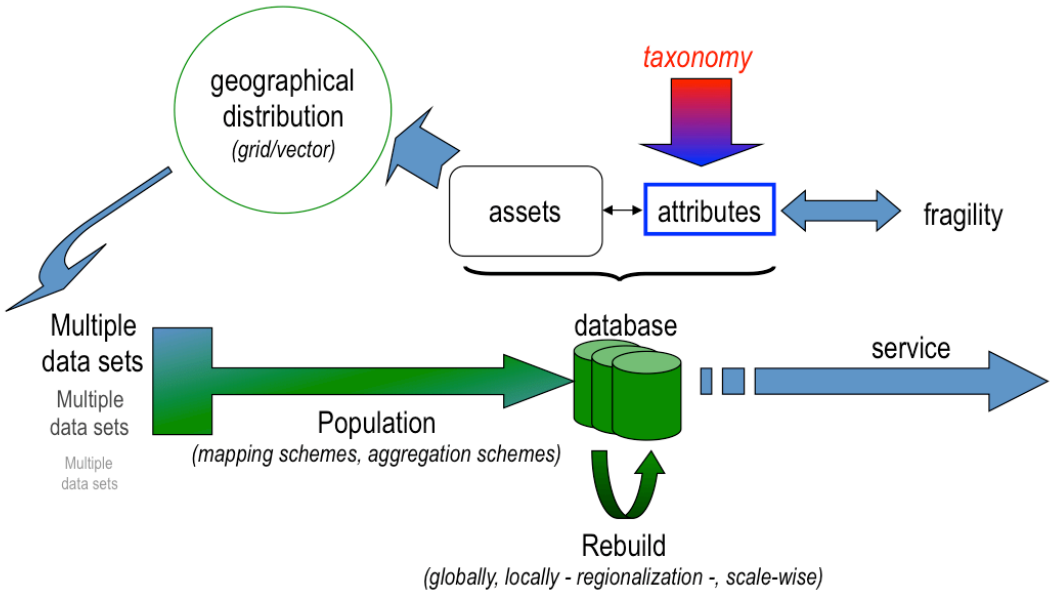

Figure 1. Graphical representation of GED inputs, outputs and features.

2. DATASETS AND MAPPING SCHEMES

GED4GEM aims to create an open global building and population inventory suitable for earthquake risk modeling. Such a database will be used in conjunction with hazard and vulnerability components to create views of risk for any region in the world. GED4GEM consortium will collect, homogenize, and build upon many existing global or regional databases at different levels to create this inventory. In order to meet the GED4GEM objectives, the GED4GEM consortium has collected global population, demographic, building, land use, and land cover datasets largely available in the public domain. Several global datasets have been identified, and a general process for integrating the data into a form suitable for loss estimation has been established as shown in Fig. 1. They have been evaluated and discussed following the SWOT (Strength, Weakness, Opportunity and Threat) analysis framework. To illustrate strengths, an assessment of how the various inventoried databases have been used to develop the GED is provided. For weaknesses, the limitations of the existing methods to exploit these data were acknowledged, as well as areas where progress needs to be made.

2.1 Global Datasets

Although there are countless datasets describing the built environment, there are very few that are directly applicable to earthquake exposure modeling without significant processing. Additionally, access to these datasets is often challenging to negotiate. Datasets for developed countries are most often available due to regular data collection for planning and commercial purposes. The challenge of the GED4GEM project is to develop a tiered database that will accommodate the best possible assumptions globally, as well as detailed data when available locally.

Vector data, or GIS data, is the most common form of spatial data analyzed in exposure and loss estimation programs. Vector data includes point, line, and polygon datasets such as building points, building footprints, postal or census zone aggregates, and critical lifelines. A typical example is the Global Administrative Areas (GADM) dataset, which is a global administrative boundaries database created to support various geo-referencing uses and census data mapping. The collaborative group overseeing the creation and development of the GADM database strives to map the most up-to-date country boundaries at all administrative boundary level, national to sub-regional, for all nations and is

open to the public for non-commercial use (http://www.gadm.org/). Another very useful dataset is the Global Rural-Urban Mapping Project (GRUMP, http://sedac.ciesin.columbia.edu/gpw/). GRUMP contains a geo-referenced framework of urban and rural areas by combining census data with satellite data. GRUMPv1 is comprised of three data products. First, GRUMPv1 provides a higher resolution gridded population data product at 30 arc-seconds, or ~1km at the equator compared to its predecessors, for 1990, 1995, and 2000. Second, GRUMPv1's urban extents dataset delineates urban areas based on NOAA's night-time lights dataset and buffered settlement centroids (where night lights are not sufficiently bright). Third, GRUMPv1 provides a point dataset of all urban areas with populations of greater than 1,000 persons, which may be downloaded in Excel, CSV, and shapefile formats (http://sedac.ciesin.columbia.edu/gpw/global.jsp).

As an alternative to vector data, remotely sensed datasets have several advantages with regard to analysing risk. Whereas, vector datasets as seen above are largely dependent on data collected from the ground, remote sensing databases are collected from the air and thus are less susceptible to bias. One example is the Global Land Cover 2000 (GLC00) dataset based on daily data from the VEGETATION sensor on-board SPOT4, data from region specific sensors, and LITES 1994-1995 for classification constrains (Bartholome and Belward, 2005). The map was derived using a supervised classification method and produced in collaboration with 30 international research groups coordinated by EC-JRC. Another example is GlobCover, which created and is able to update global land cover maps based on an automated process using the data provided by the European Space Agency ENVISAT satellite. Specifically, the system used Medium Resolution Imaging Spectrometer (MERIS) data from December 2004- June 2006 and incorporated road network data and additional layers from Australia, South America, Western Asia, Africa, India and Japan, and from GLCOO data (Bicheron *et al.*, 2008).

Finally, demographic data is commonly attribute data associated with vector data described above, but can also be single numbers supplied on a national basis. Given the absence of global building exposure databases, demographic data is essential for estimating regional exposure. For example, if the number of buildings is unknown in a given country, population estimates by regional area or postal code are likely to still exist. Combined with such statistics as number of people per household and number of households per building, it is possible to estimate the number of residential structures. Examples of such datasets are the Demographic and Health Surveys (DHS). DHS are nationally representative household surveys that provide data for a wide range of monitoring and impact evaluation indicators in the areas of population, health and nutrition. The indicators are presented in terms of national level statistics and for population subgroups such as those defined by age, education, marital status, economic status, urban/rural residence and region of the country. Household characteristics assessed include the household composition (how many people per household and their ages), educational attainment of household members, and school attendance ratios. Data on housing characteristics are also collected, including availability of electricity, water and sanitation facilities, as well type of flooring material and cooking fuel.

2.2 Mapping Schemes

Mapping schemes provide a procedure for connecting a given set of information, such as housing inventories or demographic data, to structural classifications that can be used to assess risk. Mapping schemes can be quite complex, as in the case of HAZUS, where commercial and government databases are used to assess the number of buildings associated with 28 types of occupancies, and mapping schemes are then used to distribute these buildings into structural classes that also consider quality, height, and era of construction. Mapping schemes can also be quite simple, as in the case of USGS PAGER, where on a country by country basis; the "urban" and "rural" population is distributed into the dominant structural classes in those areas. When the structural classification data are not available, the information is adopted from a compatible country. The USGS PAGER mapping schemes represent a tremendous worldwide effort, and form the basis for the initial version of the GED. These initial mapping schemes can be improved, of course, and they are currently being reviewed and modified given additional data from national housing census data supplied by UN

Habitat to be used within the GED database.

The GED consortium is also working in concert with the Inventory Data Capture Tool (IDCT) group – another project funded by the GEM Foundation, aimed at developing tools for populating the GED after its initial deployment - to support user-defined regions and user-provided mapping schemes that can be used to create GEM compatible exposure datasets representing aggregated data. The IDCT tools will allow users to create building exposure data from remotely sensed data, sample structural types with field visits, and merge these observations by developing improved mapping schemes. The outputs of these tools will be GED compatible, and depending upon the breadth of the study, will be considered for incorporation into the GED. In addition to the final dwelling and building fractions for the selected areas, the mapping schemes themselves and the original data used to compute them will be stored in the GEM system in such a way that they will be available for subsequent studies.

3. GEOGRAPHICAL LEVELS

The GED not only attempts to harmonize the best exposure datasets that are publicly available worldwide but also aims to incorporate the best practices in the creation of exposure datasets at places where such efforts are lacking. It is therefore implemented as a multi-scale collection, where spatial resolution is connected to the availability of data and its quality. The coarser resolution of the GED (referred as "Level 0") will use a simple version of the inventory available through the USGS PAGER database and based on PAGER-STR taxonomy (Jaiswal and Wald, 2008). Level 0 includes statistical information significant at the country level (and for some countries information is obtained by neighboring ones due to the lack of data sources), besides integrating the new UN-Habitat data. Significant efforts are necessary to estimate attributes at a finer spatial detail, and to develop the corresponding sub-national mapping schemes, based on recent or newly found data-sources that provide detailed inventory of building stock, aggregated information about land use classes and built-up elements on geographically smaller areas (e.g., remote sensing data and derived products) (Gamba and Herold, 2009), and other socio-economic indicators.

This leads to the development and deployment of the finer "Level 1" GED, including a 1 km x 1 km grid with as many as possible subnational housing inventory datasets from around the world. Accordingly, meaningful semi-automatic algorithms were developed to fuse the information available from different sources and facilitate geographically multi-scale and statistically consistent inventory and exposure information. Given the differences in processes of compilation of individual inventory datasets and their vintage, a simplified rating scheme was also adopted to estimate the quality and uncertainty of underlying datasets for meaningful exposure estimation.

Finally, the most refined grid level of the GED – level 2 – corresponds to aggregated data from a building-by-building inventory which might be locally available on the same 1 by 1 km grid mentioned above, instead of being computed by disaggregating aggregated data as in level 0 and level 1 data. For level 2 data, therefore, the mapping schemes involved are specific to detailed building stock information, as the data in each grid element is computed by disaggregating information available from city/county-level data. Thus, in this level the information is based on actual building stock accounts and their distribution according to a detailed taxonomy.

In all the GED levels, a common global grid is considered, and corresponds to the geographical core element of the database, stored by the grid cell's midpoint, using its latitude and longitude within a shapefile geometry. Only land areas are included; oceans and uninhabited places such as Antarctica are excluded. Basic attributes of the grid cell such as the land area, categorization as urban/rural area, and quality information related to the urban/rural classification are also stored. The exact geographic extent of the database is 180°W, 180°E, 85°N, 85°S. The grid geometry and attribute information is generated using the "Land/Geographic Unit Area Grids" and "Urban Extents Grid" datasets from the SEDAC Global Rural-Urban Mapping Project, Version 1 (GRUMPv1). Additional information about the grid cell such as the country, the administrative units – down to three levels, and the CRESTA (Catastrophe Risk Evaluating and Standardizing Target Accumulations, see https://www.cresta.org/)

zone and sub-zones it belongs to are also stored in the database. All the exposure data are either aggregated or disaggregated to the grid cell level depending on the resolution of original data.

The geometry and attribute information of countries and their administrative units down to three levels (e.g., region, province, and municipality in Italy) – are stored in the database. The parent-child relationship between these administrative units is also stored in the database using foreign key references, which are derived from the GADM database and have the ability to store changes in boundaries and names over time.

The last, most detailed element of the GED is a building-by-building vector only database, composed of a building level inventory with all the building attributes necessary to make use of detailed vulnerability functions. Such a GED level – level 3 – is currently under development and will require, for instance, a well defined data model for the building taxonomy, currently not yet well developed. Level 3 will most probably only be designed within the time frame of the GED4GEM project (scheduled to end by the end of 2013), and then populated in future updates by means of the same effort discussed above for level 2 completion.

A graphical representation of the four levels of the GED with some of their features is provided in Table 1.

Table 1. GED level of resolutions and features.

	Source	Taxonomy	Grid/Vector	Statistical significance
Level 0	LANDSCAN PAGER mapping schemes GRUMP	PAGER STR	30" (~ 1 km)	Country
Level 1	Sub-country db (UN-HABITAT, national census, regional program data sets) Urban density by RS (aggregated, < 1 km)	GEM "basic"	30" (~ 1 km)	Region/ municipality
Level 2	Local database(s)	GEM "basic"	30" (~ 1 km)	Local area
Level 3	Field survey Building database(s)	GEM "full"	vector	Single building

4. GED IMPLEMENTATION

As mentioned above, the Global Exposure Database for the Global Earthquake Model (GED4GEM) project aims to build the first globally consistent dataset of building stock exposure, specifically containing information related to structural, population and housing characteristics of the general building stock at different spatial resolutions. Due to the variety of input data utilized in the process, complexities in developing globally consistent structural and occupancy mapping schemes, and spatial and temporal interoperability requirements, the GED4GEM database design requires thorough consideration. The GED4GEM team has completed the design of the initial data model in order to accommodate available datasets as well as preserve the capability to ingest newer detailed datasets in the continued evolution of the Global Exposure Database (GED).

The GED implementation uses a database structure that is compliant to the Open Geospatial Consortium recommendations and is consistent with the OpenQuake architecture (http://openquake.org/). To improve its contents, it will be made available to the whole community backing the GEM initiative for feedback. Spatial databases with variable resolutions and structures will be linked appropriately, and standards-based tools will also be developed to support access, querying capabilities, and visualization functions.

The database schema (see Fig. 2) separates raw global data from inference, so that when inferences are updated, the raw data will remain the same. Moreover, the raw data is also saved within the database; although restricted data will not be accessible by the user. This will provide a way to rebuild one or more layers in one or more countries of the GED in a seamless and consistent way as soon as new datasets (e.g., more refined population layers) or better mapping schemes are available. Inferred attributes are to be accompanied by uncertainty and quality estimates whenever possible.

The GED data model is extremely complex and it includes a number of tables, and their specific relationships. The detail of the GED data model scheme is described in Vinay et al. (2012). The following section provides a brief description of the most important tables and features related to the database schema.

The population count of each grid cell is stored in the population table along with any available quality information about the population measure. The data model has the ability to store population data for each grid cell from multiple data sources. The population source information is stored in an additional, dedicated table. Each record (row) in the population table is associated with a grid cell and a population data source via foreign key references. Initially the database will have population data derived from the "Population Count Grids; Year 2000" dataset from the SEDAC Global Rural-Urban Mapping Project, Version 1. However, population data from other sources, such as Landscan™ (http://www.ornl.gov/sci/landscan/), or a newer version of GRUMP, could also be added.

Another table stores the day time, night time and transit population ratios, at the country level, for each combination of urban/rural categorization and occupancy classes. For example, in the beginning we have only two occupancy classes, residential and non-residential. So for each country, there will be four sets of day/night/transit population ratios: urban/residential, urban/non-residential, rural/residential, and rural/non-residential. The ratios for each country are obtained from the PAGER database (Jaiswal and Wald, 2008).

The information about a specific area, including the geometry, the study region it belongs to, and the mapping scheme applied, is stored in an aggregated building infrastructure data table. Finally, for each grid cell in a specific area within a study region, the day/night/transit population allocation ratios for each building type can be computed by multiplying the weight for that building type by the day/night/transit population ratios of that grid cell (which can be obtained from the appropriate set of population allocation ratios at the corresponding country level for the given urban/rural and occupancy category of the grid cell). Finally, using these ratios and the population data of the grid cell (see population table), the day/night/transit population counts for each building type within a grid cell are computed and stored in the aggregated building infrastructure data table along with the number of buildings and the area. Obviously, different population sources can be used to estimate the population count for each building type; the database supports such computation and the ability to store them.

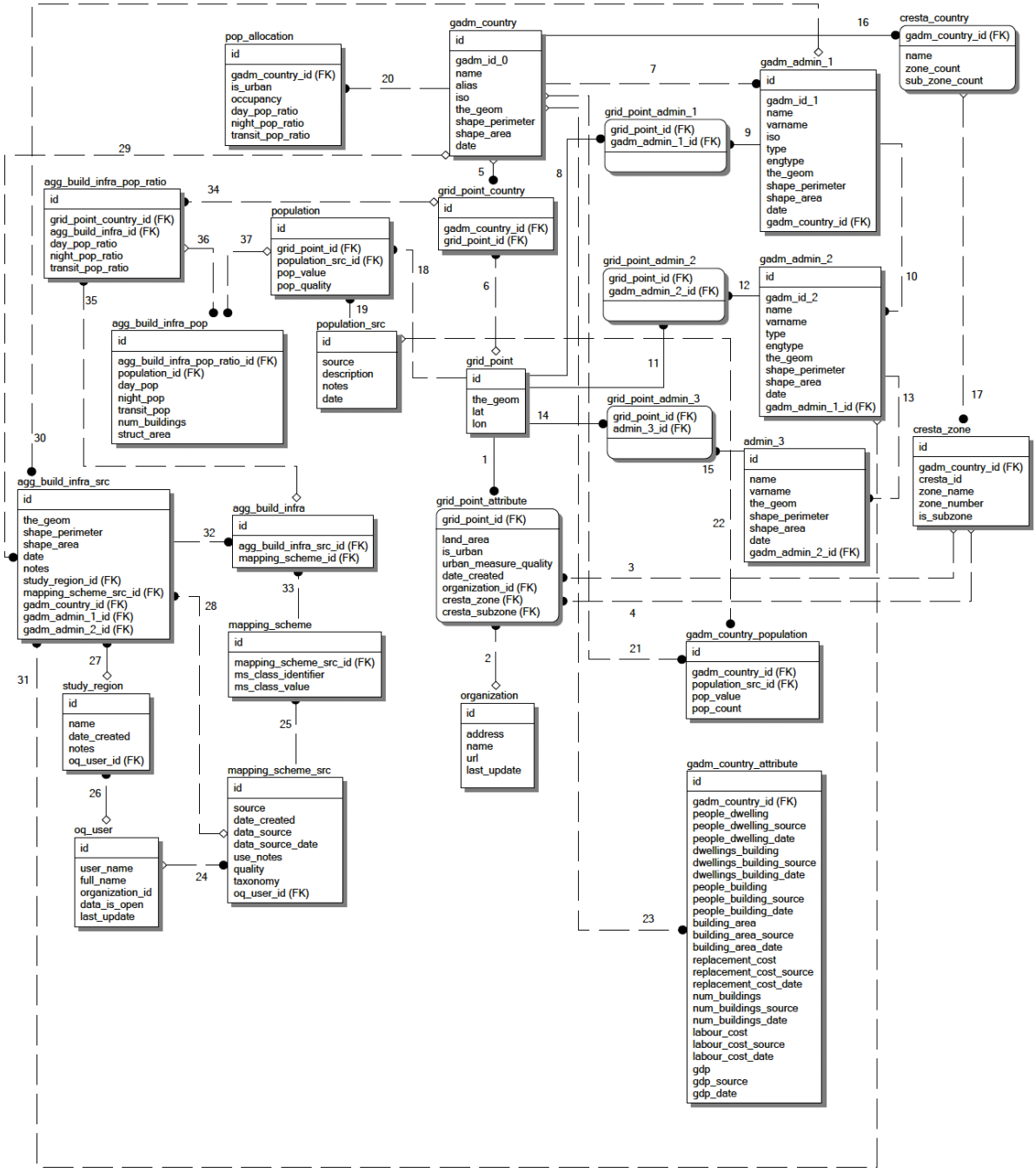

Figure 2. Current GED data model.

5. CURRENT STATUS OF GED AND WAY FORWARD

Within the challenging arena of the GEM, the GED4GEM project aims at developing a Global Exposure Database, or GED, an effort central to assessing earthquake vulnerability and risk assessment worldwide. The GED4GEM database contains grid-based earthquake exposure information derived from a variety of input datasets that are publicly available at multiple resolutions.

The GED database structure and data model have been defined. The global exposure data in terms of population counts, population distribution according to occupancy and structure-type category have been implemented at ‘Level 0’. Ongoing activities related to ‘Level 1’ include compilation of raw inventory datasets and development of country-specific structural mapping schemes. Further

improvements of 'Level 0' data are always in process. This task will be completed by the end of September 2012. The definition of the terms, database schema/structure, user manual and other guideline documents related to GED will be made available through the GEM NEXUS website (http://www.nexus.globalquakemodel.org/). The GED development guidelines and project documentation will help during outreach activities aimed at capacity building. The final version of the GED is scheduled for public release in November 2013.

AKCNOWLEDGEMENTS

The authors gratefully acknowledge the work of the entire GED4GEM team that include CIESIN, the Joint Research Centre (JRC), EUCENTRE, and the Global Urban Observatory of UN-Habitat, and the GEM Model Facility team. Financial support provided by the GEM Foundation is gratefully acknowledged. The collaboration with other Global Components of the GEM platform was essential for the design the data structure of the GED. Special thanks to Ontology and Taxonomy Team for developing the Basic Building Taxonomy that was implemented within the GED database structure.

REFERENCES

Bartholome, E. and Belward, A.S. (2005). GLC2000L: A new approach to global land cover mapping from Earth observation data. International Journal of Remote Sensing, **26**, 1959–1977.

Bicheron, P., Defourny, P., Brockmann, C., Schouten, L., Vancutsem, C., Huc, M., Bontemps, S., Leroy, M., Achard, F., Herold, M., Ranera, F., and Arino, O. (2005). GlobCover – Products description and validation report, Version 2.1, 2008 (a). Available at: (http://ionia1.esrin.esa.int/).

Dell'Acqua, F. (2009). The role of SAR sensors. In: Global Mapping of Human Settlement - Experiences, Datasets, and Prospects, P. Gamba, M. Herold (Eds), CRC Press, 309–320.

Gamba, P. and Herold M. (2009). Global Mapping of Human Settlement - Experiences, Datasets, and Prospects, CRC Press.

Jaiswal, K., and Wald, D. J. (2008). Creating a Global building inventory for earthquake loss assessment and risk management. USGS Open File Report, OF 2008-1160, pp 103. (http://pubs.usgs.gov/of/2008/1160/).

Jaiswal, K., Wald, D., and Porter, K. (2010) A Global Building Inventory for Earthquake Loss Estimation and Risk Management. Earthquake Spectra, **26: 3**, 731–748.

(NIBS and FEMA) National Institute of Building Sciences and Federal Emergency Management Agency, (2009). Multi-hazard Loss Estimation Methodology, Earthquake Model, HAZUS®MH MR4 Technical Manual, Federal Emergency Management Agency, Washington, DC.

Pesaresi, M., Gerhardinger, A., Kayitakire. F. (2008). A Robust Built-Up Area Presence Index by Anisotropic Rotation-Invariant Textural Measure. *IEEE Journal of Selected Topics in Earth Observations and Applied Remote Sensing*, **1: 3,** 180–192.

Polli, D., Dell'Acqua, F. and Gamba P. (2009). First Steps Towards a Framework for Earth Observation (EO) Based Seismic Vulnerability Evaluation. *Environmental Semeiotics*, **2: 1**, 16–30.

Schneiderbauer, S. (2007). Risk and Vulnerability to Natural Disasters - from Broad View to Focused Perspective, Ph.D Thesis, Free University of Berlin.

Vinay, S., Chen, R., Becker, M., Huyck, C., Hu, Z., Cavalca, D., Goldoni, E., Gamba, P., and Jaiswal K. (2012). Description of the Global Exposure Database Schema, Report produced in the context of the Global Exposure Database for the Global Earthquake Model (GED4GEM), GEM Foundation, Pavia, Italy.