A Way Ahead?

Suggestions to initiate discussion
toward a consensus goal and plan

GEBCO: Authoritative?

GEBCO wants to be “Authoritative”.

[words are from GEBCO's December 2006 brochure]

But are we?

GEBCO's existence under international agreement and cooperation does not imply authority or quality in the eyes of today's user.

Who we are doesn't matter.

What we do and build will earn our reputation.

Our old way didn't achieve this

Old GEBCO: wait for others to produce products (grids, contours, etc.) and offer them to us, after which we stitch them into our patchwork quilt. This fails because

- We have to negotiate terms of use each time
- We have to edge match
- ***Quality and update schedule are not in our control***
- We are perpetually out of date:
  - Updates happen only when others build them;
  - The user can obtain the new information faster by getting it directly from the originator.

How to be authoritative and best available

If we are ***open*** about our data ***sources*** and the ***methods*** we choose for deconflicting, harmonizing, and interpolating among heterogeneous sources, and…

If it is ***evident*** that our products are built from the most ***comprehensive*** diversity of sources, using what international experts agree are ***best practices***, …

Then we will earn the labels "authoritative" and "best available" in today's open source world.

An example of open & evident

The "Smith and Sandwell" products make evident where they rely on soundings and where they interpolate. The interpolation schemes are published in a peer-reviewed journal. The user can judge whether the distribution of control soundings is comprehensive and reliable. [See poster by Binder.]

It is those features, not the altimetry or the authors' names, that make the product attractive.

In Aguascalientes in 2005, SCDB XXI agreed to work toward building a GEBCO bathymetry in a similar fashion.

We need a Work Plan and Schedule to get there.

First steps taken

We have started to tackle direct ingest of *x,y,z* data.

The GEBCO gridders found that grids could not always be built from contours alone, but sometimes had to have *x,y,z* values added.

With the ENC *x,y,z* harvesting project, we are venturing into building the shallow water parts of the grid directly from sounding data. [See poster and presentations by Weatherall and by Pharaoh.]

We drafted new Terms of Reference (October 2006).

The Sub-Committee aims to:

1. Maintain and improve GEBCO products and supporting data;
2. Monitor developments in data availability and relevant technology [to] maintain the excellence of GEBCO products;
3. Provide advice on the scientific and technical aspects of bathymetric mapping;
4. Encourage and facilitate the location, acquisition and exchange of data supporting bathymetric mapping;
5. Produce products that are easily applied to other ocean sciences;
6. Establish, nurture, and/or disband working groups, as needed, to carry out specific tasks or product developments that relate to the technical advance of the GEBCO project;
7. Work with SCUFN on matters of joint interest.

From the October 2006 proposed revisions to SCDB's terms of reference.

1, Maintain and improve GEBCO products and supporting data…

… such as, but not limited to:

1. A global bathymetric grid;
2. The GEBCO Digital Atlas;

[I think this language (October 2006 proposed ToR) meant to highlight both the grid and the contours as distinct products or responsibilities. In my view, it is not useful to separate the grid from the GDA.]

We must first have a bathymetric <u>model</u>, from which contours or grids may be produced.

On top of the model there may be a user interface, e.g. the explore / display / output functions of the GDA.

This presentation looks at how we might build / maintain / improve the General Bathymetric <u>Model</u> of the Ocean.

Why build our own model?

So that we control the quality and update schedule of our products.

We choose, by consensus of international experts:

- Quality control procedures for input data
- Deconfliction, generalization, and interpolation methods.

We make our strengths and weaknesses evident:

- We show where we have source data
- We show what our sources & algorithms are
- We [eventually] build a model of the quality or uncertainty in our bathymetric model.

1, Maintain and improve GEBCO products and supporting data…

… such as, but not limited to:

1. A global bathymetric grid;
2. The GEBCO Digital Atlas;
3. Databases of soundings, shorelines, land elevations, remotely sensed and other data, generalized to a useful working scale, as may facilitate update of GEBCO products and maintenance of product quality. [Again from October 2006 proposed ToR.]

Item 3 says that ***we take control of the data behind our products***, and that ***we maintain those data*** in a form that allows us to readily update our products. This need not mean "raw" data but data QCd and thinned or generalized into some form appropriate to our current & future work.

Why build our own database?

So that we can quickly & easily update our model.

In my experience, the biggest effort is data ingest:

- Finding the data in diverse sources worldwide
- Getting it into a common format
- Quality-controlling the data

Once we have made this effort, we must keep the result, so we don't have to do it over again to update.

But we also need to be forward-looking as we ingest data, so that we get what we need to build the future model, not just the present model. (Data ingest efforts now must anticipate building not only a 1' grid but also a future grid with a smaller or variable mesh.)

Source data policy questions

Should the GEBCO model of global bathymetry be built ***only*** from “publicly available” data? [Quote from the December 2006 brochure. Was this issue decided?]

What if someone wants to contribute data in a restricted or proprietary way? (e.g. "you can use it to update the GEBCO 1 minute grid but don't share the raw data", or "you can use my derived grid but not my underlying data") If we use such data, how do we protect them?

Are people unlikely to give us raw data?

Are we unlikely to be able to handle storing it?

Does GEBCO need a written data-sharing policy?

One database suggestion

Raw data could be distributed in archives maintained by the originators of the data, not GEBCO.

GEBCO would maintain "abstracts" of QC'ed data at a working scale (e.g. 0.1 arc minute, about 200 meters ?), from which our bathymetric model products can be built.
[The GEBCO V edition external product was a 1:10M map but the internal working scale for compilation sheets was 1:1M. This is a similar idea.]

An "abstract" reduces each source to a summary of statistics for each geographical tile or block:

- Minimum, mean, median, and maximum depths
- Uncertainty estimates in the above

We could build the abstract from raw data ourselves, or let contributors send it (we could supply software).

Techie fine point on tile size

If an "abstract" summarizes a "tile" of latitude and longitude smaller than 0.25 arc-minute, then the number of such tiles on Earth overflows a 32-bit integer.

Some thought should go into how to number or identify tiles so that they can be readily indexed and sorted.

The "MOA" example (a few slides later) supports tiles which are 1 or 2 arc-minute Mercator projection pixels (old Smith and Sandwell style) and also the 30-arc-second lat,lon pixels used in GTOPO30 / GLOBE / SRTM30. All of these can be indexed by a 32-bit value. If GEBCO uses a 0.1' tile, things are more complicated.

Compilation from abstracted data

The abstracts of each individual source are brought together to build an aggregate of all sources covering the same tile area.

Deconflicting and estimating combined statistics is done at this step, along with further QC.

For example, one may find that a particular source is always the shallowest or deepest in each box it crosses, and this may cast suspicion on the source.

The output of this step is an aggregate of all sources, from which the model is built.

Model building

From the aggregate of all abstracted data we build a model by some process of interpolation and perhaps also smoothing. Possible processes include:

- GMT, used by the GEBCO gridders
- Altimetry-guided methods (Smith & Sandwell, French group, or others [I'm experimenting with some new ones])
- Statistical methods (aka kriging, objective assimilation, collocation, etc.)

The method could also yield an uncertainty estimate.

Output

The model can generate

- Grids of depth/elevation
- Contours of depth/elevation
- Confidence or uncertainty limits on depth/elevation
- Grids indicating control locations and lack of control
- Grids of source identifiers [example: Kim's poster]

Example: the MOA project

On-going, unfinished, draft interim grid available now.
(One arc minute Mercator format so far, see posters.)

Collaborators: NOAA, NGA, US Naval Oceanographic Office, Scripps, BODC, IHB, and many data contributors and colleagues. ("MOA" applies to the U.S. government agencies. I need a more inclusive name for this effort.)

Data manipulation software is "home-grown" and has not been GIS-enabled. As above, "abstracting" is supported only on a few rather large (0.5 to 2 minute) tile sizes.

This may or may not be a model for GEBCO.

MOA, 2: distributed effort

Collaborators work in parallel on

- Data ingest and input data QC
- Data flow
- Algorithm development and test
- Output model QC

as the interests and resources of each allow.

To date, everything is sitting in one computer, and collaborators log in remotely. Things are not as distributed as they could be yet.

MOA, 3: sharing

Data are shared and open to all participants at all stages (ingest, QC'ed, "abstracted", and modeled).
Software is shared and open to participants.
Participants can thus duplicate the effort in their own shops and add their own proprietary data if they wish.
The main product should be open from end to end.

Sharing is accomplished through a password protected account on a designated computer system (not distributed across systems).
The U.S. NGA is also contributing proprietary inputs that control the model yet are "hidden" on output. This has slowed down the sharing of the project outside the U.S.

MOA, 4: data ingest

Single-beam cruise files and x,y,z lists are currently ingestible. ENC / DNC soundings and multibeam swaths are ingested as x,y,z. Other ENC/DNC layers are not currently ingestible. Data in the form of "bounds" (e.g. "below sea level but not deeper than 50 m") are not yet used, but easily could be.

Each source is given a unique identifing number.

Data are translated to a common file format.

We have not dealt with vertical datum yet.

MOA, 5: common format

Each translated source file has this information, one record for each sounding or point:

- time, or sequence number in the file;
- latitude;
- longitude;
- depth;
- depth uncertainty;
- position uncertainty;
- depth according to previous bathymetry model.

The depth and position uncertainty may be left unknown. The depth in the previous model facilitates initial QC by furnishing a sanity check. It is not given weight "downstream" (though it could be).

MOA 6: abstracting & aggregating

/Users/walter/Desktop/New_Predict/src/BMsrc/blockstat.h

```
/* blockstat.h

Defines some structures and constants used by the blockstat
subroutines and main programs.

W H F Smith, 9 Jan 2007, version 1.  */

struct BM_DATA_IN {
    int index;      /* identifier for BM pixel  */
    int source_id;  /* unique file number for the source of this point value */
    short int z;    /* a depth or elevation value, neg below sea level */
    short int s;    /* combined H and V uncertainty in this z value */
};
```

Each data point is reduced to a tile index, source ID, depth and error est.

The *x,y* error is scaled by the current model's seafloor slope and combined w/ *z* error. If errors are blank, default values are assigned based on age and tech.

```
struct BM_DATA_OUT {
    int index;  /* range from 0 up to 933 120 000 - 1 for srtm30  */
    int id_min; /* source ID associated with minimum (deepest) z value */
    int id_med; /* source ID associated with median value */
    int id_max; /* source ID associated with maximum (shallowest) z value */
    short int z_min;    /* minimum (deepest) z value */
    short int z_max;    /* maximum (shallowest) z value */
    short int z_mean;   /* mean z value */
    short int z_std;    /* standard deviation of z values */
    short int z_median; /* median z value in this block  */
    short int z_spread; /* inter-quartile range of z in this block  */
};
```

Error-weighted abstract aggregates all points in a tile

MOA 7: model production

For now, using old method (Smith & Sandwell 1997).

This assumes all data are equally certain.

We are experimenting with new methods.

We hope to develop a method that will propagate error estimates and furnish a map of uncertainty. [The poster by Marks is a step toward an error budget for the data ingest side.]

For now, the only output quality indicator is the old one: grid points are flagged as being either constrained or not, based on whether or not the corresponding tile has aggregated control or is empty.

Examples are on the posters.

GEBCO vis-à-vis the MOA

Option 1: GEBCO adopts this project as it currently is.

Option 2: GEBCO adopts aspects of the project and improves or reshapes them to better suit GEBCO.

Note: These two are not mutually exclusive.

I think this was our intent in Aguascalientes.

I hope we build a Work Plan aimed at both.

Option 3: GEBCO does something else altogether.

But in that case I have no idea what to do.

Option 1: GEBCO adopts MOA project as it currently is.

We agreed in 2005 to do essentially this, though the details were not then clear. The delay was caused by the U.S. federal agencies slowly working out their data sharing agreement. Now I think we are ready to go. GEBCO's ENC xyz harvest is ingesting into the MOA, as a first step.

Strength: This project is underway.

Weakness: So far, things have been held in U.S. and not fully opened to GEBCO yet, due to some proprietary U.S. data in the project. (The use of altimetry is no longer viewed as a weakness, right?)

Opportunity: GEBCO can make the product internationally comprehensive in data and QC.

Threat: The MOA product will emerge soon and will be a competitor to GEBCO unless GEBCO co-opts it.

Option 2: GEBCO adopts, improves, reshapes MOA

Strength & Opportunity: There is enormous room for improvement, but the skeleton may serve as a guide.

With openly shared data and software, each IBC or participating institution could build its own "GEBCO plus" model, by using common algorithms and adding its own in-house proprietary data to the common GEBCO data.

Weakness & Threat: If GEBCO does not also do Option 1 simultaneously, it risks falling too far behind. The project will donate its data to GEBCO, but if GEBCO does not like what the project has done with the data and metadata, GEBCO might end up having to re-acquire data.

Work to do under Options 1 or 1&2

Data ingest: All of you please look at the Binder *et al*. poster showing data obtained so far, note what data we do not yet have, and help us get that data. Write your email address by your remarks directly on the poster.

Model QC: All of you please examine the model, looking at the posters here and also by downloading and examining it, comparing it against your proprietary data. Again, write on the posters.

Data management: Build a "mirror site" where a GEBCO version of the data can grow, so we can disengage from the MOA site that holds U.S. proprietary data. Access to the GEBCO site can then be fully opened. Data donors will feel that they are giving to an international project. Contribute your expertise in distributed and GIS-enabled databases to improve how we build and maintain the data that underpin the model.

Algorithm improvement: Help us figure out how to do it better. We want to get to lat,lon from Mercator; to handle error-weighted and error propagation schemes; to tune parameters & try new things. Antarctica has some particular issues (depths in cavities under permanent ice).

Data management & policy

Are policy issues decided? Should GEBCO build a common global model based only on public data? Who will have access? How will we protect the database?

Decide the tile size for internal workings and abstractions, looking ahead to finer or variable grids.

Decide the right items to go into the database and the abstractions. Figure out how to link to metadata.

Decide, design and build using the full breadth of GEBCO expertise.

Do we need an ad hoc working group to do just this?
Or is SCDB small enough that we can do it all together?

Proposed work timetable

At this meeting:

- Decide on Option 1, or 2, or 1&2, or 3.
- Comment on MOA posters.
- Adopt policies and work plans as appropriate.

Now or very soon (e.g. before 1 January 2008):

- Freeze the GDA and Grid. Stop updating. Put out the version or updates we have, as appropriate. [Call it the “2007” version?]
- Find new data to add to new model, QC new model.

During 2008:

- Bring out interim new model grid.
- Build "mirror site" or other GEBCO data base.

During 2009:

- Add to database.
- Revise model grid.
- Build GDA or other products derived from model grid.

By 2010, a new product suite is released.

Let's not forget:

GEBCO is not just about building a bathymetric model; it is more than just its grid. Additional things include:

- The GUI functions of the GDA.
- The Nippon Foundation training program.
- The relationships through IHO and IOC.

I have focused on building the General Bathymetric Model of the Oceans. I haven't addressed the above. Any work plan should include these as well. All of these will evolve as our way of working changes.