TOO MUCH DATA OR NOT ?

G. Ducher, I.G.N. (F), Paris.

I. INTRODUCTION

It is usually said that we are entering an era of huge production of data. It is often added that the risk is to be overwhelmed with the vast amount of data that is collected everyday throughout the world.

It can be of interest to try to assess an order of magnitude of this amount, using some examples related to photogrammetry and remote sensing and applied to topographic mapping and geographic information systems (GIS). So a review of some recent aerial and space activities and trends will be given, quoting some figures about the volume of data which is collected and can be considered as converted into geographic information.

As a conclusion, it will be perhaps easier to answer whether too much data is available or not, if any answer can definitely be made to such a challenging question.

II. AERIAL PHOTOGRAPHY

It is very difficult to draw up a comprehensive list of the worldwide situation as regards available aerial coverages and to provide a complete inventory [1].

As regards France, I.G.N. has taken systematic stereoscopic aerial photographs since 1945, achieving complete renewal of coverage every fifth year, on average. About 100 000 km2 are re-flown at 1:30 000 scale by the IGN Aerial Surveys Department, and additional pictures taken at larger scale, (1:15 000 to 1:20 000), on panchromatic film doubled on black and white, color infra-red or true colour film, over 30 to 40 000 km2, for the National Forest Inventory, urban management and other planning purposes. As a result, 35 000 to 55 000 pictures are taken each year over the French metropolitan area, while a total of 50 000 to 80 000 pictures is obtained, including surveys over French overseas territories and foreign countries, at their request ; on average, the annual production can therefore be stated to 70 000 frames.

It can be interesting to assess how many bits this production would represent, if it were taken with an area array camera or digitally transferred from film on to longer life-times bases, to comply with the requirements of a better storage or to be computer processed. Assuming that a 20 microns slit width should be used for scanning the pictures, which seems quite a convenient slit on average, and that each pixel should be coded into 256 grey levels, i-e 8 bits, or 1 byte, each 23 cm x 23 cm format photograph would give rise to about 10^8 pixels and 10^8 bytes. The I.G.N. (F) annual production of aerial photographs would then represent 7×10^3 gigabytes, i-e 7 terabytes. At present nearly 4,000,000 aerial photographs are stored at the I.G.N. (F) photographic library [2], which would represent 4×10^{14} bytes, i-e 400 terabytes. This figure gives an idea of the amount of raw data which is accumulating in store and of the dramatic problems to be solved the day when, if any, these archives should be digitized and transferred onto tapes or disks.

It can be assumed that more than 5 million km2 of the earth's surface are recorded photogrammetrically each year (USSR and China not included), [1], i-e 50 times the area covered over France, and that the collection of aerial photographs stored in different countries probably consists of over 20 or 30 million frames, taking into account the archives of some photo-libraries such as those from Quebec (1,400,000)

U.S.A. (6,000,000), U.K. (4,000,000), F.R.G. (2,000,000) [2]. This collection would represent perhaps 3×10^{15} bytes, if not 10^{16} bytes, which are stored at the present time in analog form throughout the world, with an annual increase of some 10^{14} bytes.

III. THE USE OF AERIAL PHOTOGRAPHS

Aerial photographs are considered as a primary source of geographic information by a wide range of users, amongst whom are technicians, managers of rural spaces, urban planners, designers of projects, experts in major natural hazards, geologists, photo-interpreters, geographers, students and the general public as well. They generally use duplicates, paper prints, enlargements, mosaics, in analog form as a whole [2]. For example, each year the I.G.N.-Photothèque has a total production amounting to about 60,000 duplicates, 250 to 300,000 black and white paper prints, 20,000 black and white enlargements and 10,000 true-colour enlargements. Similar productions are reportedly achieved in other countries where a free-access policy is practised.

Most of these users aim at specific purposes and need only a very small part of all the raw data which are delivered to them and which they normally use in analog form. A few others are deriving digital geographic information from those aerial surveys, as I.G.N. (F) does for example.

In this way, an altimetric data base has just been established at the I.G.N. (F), by digitizing the contours from the 1:25 000 scale basic map ; this data base covers the entire area of France and consists of about 180 million points whose three coordinates are stored and reach about 1,5 Giga bytes. Considering that the contours have been stereoplotted from aerial surveys and that a full stereoscopic coverage of the French territory requires about 40 000 photographs at 1:30 000 scale, digitizing these photographs would have generated $40,000 \times 10^8$ bytes. Comparing the information stored in this altimetric database with the original raw data, a ratio of one byte over 2 600 bytes appears between both of them, which is a very poor one ! All the more as ancillary data, required for structuring and handling the base itself are included in these figures.

At the present time, similar aerial surveys are starting to be digitally stereoplotted at the I.G.N. (F) in a long term programme, to establish a topographic database due to replace the line base map series at 1:25 000 scale [3], and which will comprise some 20 to 25 gigabytes, including attributes and other data derived from field completion. Due to the fact that many planimetric features are extracted from the aerial photographs, for topographic mapping, the ratio between the number of bytes of this topographic database and that of the original aerial survey, if digitized, is far better this time, 1 over 200. But if we consider that establishing this database will take 30 years, a period during which France will be totally reflown 6 times, in order to collect data for map revision and other users, this ratio should decrease down to six times less, i-e 1 over 1 200.

It can be concluded that aerial surveys deliver a wealth of information out of which only a very small part is actually used and transferred into geographic information.

It could be interesting to assess which ratio will be obtained for the digital products which Ordnance Survey is delivering, and which will consist respectively of 6 gigabytes as regards the database established from maps at 1:50 000 and maybe 80 to 100 gigabytes as regards that from base scale maps (1:1 250 and 1:2 500), once these products will be achieved throughout the United Kingdom.

IV. SPACE IMAGERY

Over 2 million LANDSAT scenes have been collected from the beginning of the system ; three years ago 660,000 MSS scenes and 26,000 TM scenes were reportedly stored at the EROS Data Center and more than 1,270,000 abroad [5]. The volume of data collected is a linear function of time and is used to increasing annually by 100 000 scenes. The area covered can be estimated at about 40 Gkm2, representing 300 times the continents, with an annual increase of 3 Gkm2, i-e 23 times the continents. Of course many scenes are cloud covered, so that some countries are not yet entirely covered.

Each MSS scene comprising 30 M bytes and each TM scene 250 M bytes, the resulting amount of data collected by LANDSAT should reach 60 to 70 terabytes, which is surprisingly 6 to 7 times less than that of the only aerial photographs stored at the I.G.N. (F) over France, if they were digitized ! This comparison shows that photographs still remain an efficient storage base and/or that space is not yet really overwhelming everybody with data ! Of course a problem is created for storing and maintaining the related magnetic tapes, considering that hundred thousand 6 251 bpi CCT are required,holding 160 M bytes each and having an approximate lifetime of 3 to 5 years.

However efficient a photograph is for storing data, it is surprising that but few photographs have been taken from space. During the Metric Camera mission, ESA acquired 1,054 frames and during the LFC mission, NASA acquired 2,100 frames. It is probably the same number of photographs which USSR takes from each Soyuz-Saliut experiment, reaching a total of maybe 60 to 70 000 photographs, covering probably more than 100 Mkm2 with very good quality products. As a result, the total number of space photographs appears very limited, in comparison with that of aerial photographs.

And now, what about SPOT ? The number of SPOT images acquired by the system is regularly increasing, as a function of time and of the number of receiving stations. Over 1,200,000 scenes have been collected up to now,which represents 30 times the land area of the earth, but only one tenth of the area covered by LANDSAT. The annual increase reaches 450 000 images from SPOT, which is equivalent to 10 times the continents. Each SPOT scene comprising from 27 to 100 Mbytes, i-e 50 on average, the number of data collected should reach 60 terabytes, i-e a total quite similar to that from LANDSAT, but obtained in 3 years instead of 17, and which is still 6 to 7 times less than the aerial photograph equivalent data archived at the I.G.N.-Photothèque. However the annual increase of SPOT data is as a whole higher than that of the I.G.N. (F) aerial coverage, 22 terabytes versus 7 terabytes, which should not be surprising ; but removing the cloud covered SPOT scenes, the result should not be so different.

However the volume of SPOT data converted into geographic information is not so large, since the number of space images purchased can be assessed as being 10 %, due to cloud cover problems, lack of funds and to the SPOT-Image policy of obtaining a worldwide coverage, even before receiving orders for each scene. Moreover only 25 % are processed aiming at cartographic purposes, which reduces to less than 2,5 % the number of space images used for delivering geographic information. Considering that digital space imagery is mostly efficient for image mapping and for the establishment of geocoded products, rather than line maps, and considering the development of these products, the ratio between the number of bytes available from space and that actually used for mapping could reach 1 or 2 % in a near future, which is a better result than the ratio of 1:200 mentionned before, when using analog aerial photography, though it still remains of the same order of magnitude. However nobody can be satisfied with such a poor figure, all the more as among the 90 per cent SPOT images which remain in archives, another 10 per cent at least have the information

required for complying with the needs of many countries [3].Technical, administrative and financial solutions have still to be determined for a wider use of space imagery.

V. PROBLEMS OF COST

As regards the cost of a byte captured from aerial surveys, it depends upon the scale, location and size of the survey. Studying some results internally reported by the I.G.N. aerial surveys department, it can be inferred that this cost is ranging from about 2×10^{-6} FF a byte taken on a regular basis, over France at 1:15 000 scale, to about 10^{-5} FF, for overseas surveys. On average this cost can be assessed at 5×10^{-6} FF a byte, which looks very cheap (but what could be done using one single byte, or even with 1 FF of bytes covering a few cm2 on a paper print ?).

As regards the cost of a byte from SPOT imagery, it is independant of the location and size of the survey (unless the corresponding acreage be very small !). Considering that a SPOT-P scene is listed about 10 000 FF and contains 27 Megabytes if vertical, a byte costs 0.37×10^{-3} FF. Comparing this cost to that of a byte from aerial survey, which is 5×10^{-6} FF on average, with a maximum of 10^{-5} FF, (after digitizing the photographs, but without including the cost for digitization) a byte from SPOT appears to be more expensive than a byte from aerial surveys, in a ratio ranging from 35 to 75 times more. In fact, for a given acreage, since the number of SPOT pixels required for covering the area is 280 times less than using digitized aerial photographs at 1:30 000 scale, it is on the reverse from 4 to 8 times cheaper to acquire SPOT data instead of aerial photography at such a scale.

In fact, aerial photography would better compare with SPOT when taken at smaller scale. For example, Ottoson (Sweden) quoted a cost of US ¢ 1 per 1.1 km2 for obtaining a photographic cover of Sweden at 1:150 000 scale [8]. SPOT data, involving 12 000 pixels for each 1.1 km2, at 0.37×10^{-3} FF one pixel, would cost 4.5 FF, still remaining slightly cheaper than aerial survey at US ¢ 1 !

Considering the various assessments which demonstrated the suitability of SPOT data for contouring at 20 m vertical interval [4], with additional intermediate contours at 10 m in low-relief terrain, it can be stated that the SPOT altimetric content is only twice less than that of aerial photographs, which are usually stereoplotted into contours at 5 m vertical interval, and 10 m interval in steep slopes. So that SPOT is very cost-effective for the establishment of contour plots, all the more as the number of ground control points required for a block adjustment can be largely reduced using a SPOT space triangulation, maybe down to 50 times less, that no field completion nor any other manual involvement is required for such products and that the implementation of automatic digital correlation will reduce still more considerably the time and manpower required for obtaining DTM and contours. Finally, SPOT is at least 3 to 4 times more efficient than aerial photography, comparing the amount of information on contours to the cost of acquisition, and probably 5 to 10 times more, if the reduction of time obtained in SPOT stereoplotting, even manually, is taken into account. Similar results are reported by hydrographers, over shallow water.

SPOT is therefore mostly appropriate to the establishment of DTM over wide areas still devoid of base maps. Using SPOT saves time and money ; the fact that it saves also pixels in a ratio of 1 over 100 or 140, in comparison with digitized aerial photography, highly contributes to such an efficiency.

More detailed studies are underway in order to obtain a better knowledge of the economic efficiency of geographic information and other products derived from space data. Specifications should be clearly established, for a wide range of standard products, specially designed to overcome the contradiction between the urgent needs of

many countries and their difficulties in funding mapping projects [3]. Even though some mapping phases are 5 or 10 times less expensive using SPOT than an aircraft, difficulties are encountered in identifying small features, involving more time consuming and expensive operations. Moreover the suitability of space data to cover much wider areas in shorter delivery times, could be an incentive to increasing the total mapping expenses, the reduction in cost per km2 being overpassed by the increase in coverage, which is very challenging for many developing countries !

However, the SPOT Image Company, with a turnover of 100 MF last year, is on the verge to balance its budget and work on a full cost recovery basis. It means that space mapping is promised to an expanding activity. The creation of some new French value-added companies, dedicated to SPOT products, witnesses for a real market [6]. As regards I.G.N. (F), a new department, called I.G.N.-Espace, specially devoted to mapping from SPOT data, has just been created at Toulouse. It aims at producing mosaics, derived products and DTM.

VI. OTHER DATA

Planetary missions, which started 30 years ago, have resulted in a collection of over 500 000 pictures of planets and satellites, stored in dedicated Space Image Centers, such as JPL, the Smithsonian Institution, or the French Orsay University. Maps have been carried out, and some of them digitized, as for instance those at 1:2 000 000 scale, over Mars, scanned by I.G.N. (F). Their number will undoubtedly increase in the years to come, as many projects are being prepared, such as Mars-Observer, giving rise to an unlimited source of data.

ESA earth observation projects, including very promising new sensors dedicated to agricultural and vegetation species mapping, some of them fitted with a 5 m IFOV and an on-track stereo-imaging capability, are expected to deliver additional data flows of up to 0.86 terabytes a day, from HIRIS (high resolution imaging spectrometer) or 1.76 terabytes a day, from a SAR. The volume of data which will be currently acquired by satellite could be soon many times larger than now, requiring appropriate storage facilities and new processing centers.

Other land related data which include a reference to a two or three dimensional position as one of their basic attributes, should be considered as integral part of GIS/LIS (geographic / land information system).

Urban planners, experts from local communities are developing systems for handling lots of data. Some are already operating, such as ICOREM at Marseilles (F), consisting of hundreds of Megabytes. The Los-Angeles county foresees to manage no less than 1 terabyte in a near future. Cadastral surveys consist of hundred thousand map sheets relevant, for example in France, to 35 million landowners and 97 million parcels ; they would represent dozen gigabytes if they were digitized, in addition to some 40 gigabytes already stored from the Land Register. Another class of data comprises data resulting from all kind of distribution networks. The United Kingdom reported that GIS should integrate some 1,650,000 km of underground lines and mains, 300,000 km of overhead lines and 100 million connections of public utilities to customers. How many gigabytes would evolve from these surveys ? In fact, a slight increase of 10 % is noticed in the annual budget devoted to GIS/LIS by local communities, which is undoubtedly higher than the increase of expenses paid by national agencies for conventional topographic base mapping.

Aerial and space data is only a small part of all the data collected without arguing, all around.

Terrestrial and oblique photographs, taken for general inventories of historical

monuments and buildings, should be also taken into account in GIS. Million photographs have reportedly been taken over architectural and archeological sites in UK, and in France, where only one tenth of the monuments worthy of notice have actually been surveyed and 2 millions resulting pictures stored in archive areas, in analog format, out of which only a tiny percentage has been used as a vehicle for measurements. Similar results are reported by CIPA in other countries.

A gap is again separating the huge amount of data recorded on film base from the small quantity of data digitally transformed into geographic information.

VII. CONCLUSION

It should not be concluded whether we have too much data or not, before a complete review be carried out to determine the exact number and characteristics of the aerial and space data required for meeting the needs in topographic and thematic surveying and mapping, throughout the world, whether the projects may be funded or not.

Are we sure we have enough data for a global approach of the worldwide problems ? Are we providing enough up-to-date new products suitable for monitoring,modelling and predicting our environment ? Real time automated mapping could be achieved for example for polar sea ice floe distribution using SPOT data and future ERS-1 radar data, if an international agency was funded aiming at this purpose. Many other applications could be similarly considered for surveying changes in vegetation and land resource evaluation.

More quickly and better information will be needed in the future for new applications which are not yet fully operational but should not be overlooked today. Former data, stored in archive, will be of paramount importance later on, when required for a better knowledge of evolutions, then paying for the care taken in storing data which could seem useless today.

A static world would not need further data, nor improvements in the present types of space data. Remote sensing, devoted to a real and dynamic world, is on the reverse a very disturbing matter, still requiring many efforts for being fully operating. We have seen that data from space can already solve many problems efficiently, even keeping the number of data required to a lower value than that generally encountered using aerial or terrestrial data, in which much more equivalent pixels are discarded or stored without being used.

Besides,is not a similar waste noticed in many human activities ? How many books are read, how many roads or trains are empty ? Are not much more cars parked at any time than those running, probably in a ratio that can be estimed to 1:50, on average, (everybody driving 10 000 km each year at 60 km/h , or half an hour per day), which is not so different from the ratio between data received from space and those actually processed into geographic information ! And nature is still much more prodigal of its wealth than men !

The fact is that long-established mapping habits have to be rethought and modified for taking full advantage of space data. The areas covered are much more extended, the scale smaller, the imagery is in digital format, the amount of data to be merged requires more skilled technicians operating on more expensive and sophisticated computer-assisted workstations, and new products are to be designed and proposed for a global use and not as map substitutes. But, in the long run, the trend will bring everybody to be less wasteful and more efficient using space data than aerial surveys, and we have seen some examples showing that it is sometimes already achieved to-day.

The only way to find most suitable images for a given application is to have a large choice of data available from different systems, operating on a more or less complementary basis. The relevant incoming flood of data will bring agencies to implement advanced systems for automated data processing, automatic feature extraction and correlation ; they will have to make heavy investments in highly skilled engineers, interpreters, and training experts. But the number of available space data is not finally so impressive that it could not be tackled with.

Cooperation is a prerequisite for developing the use of space data, transferring data from different sources into GIS/LIS. A main problem should be to review who holds some kind of data, and how a free access is possible to whoever feels that this data can be of interest for him. Tentative figures place the benefit to cost ratio of sharing spatial data between organizations at around four to one. Everyone will be compelled to share geographic information and cooperate more intensively. The danger will no longer lie in an excess of data but in the risk that insufficient or untimely use be made of the irreplaceable treasure constituted by updated aerospace data.

Data creates data. Each agency using output data from other agencies, carries out new products and processed data which are used as input by other users. A wide network of inter-related data is being progressively established. New technology offers new tools. Field, aerial and space data-capture systems procure data streams which can now result in new plots of a quality and in a time that a human cartographer cannot match [7], and are quite appropriate to specific purposes.

Would it not be a solution that state-owned agencies be responsible for some standard data pre-processing and disseminate new user-oriented geocoded data on a regular basis ? Should they not be partly subsidized or encouraged to proceed in this way ?

Should not a less commercial open sky policy be set up, that makes space data more easily available and helps reducing the gap between developing countries and industrialized ones, which is not exactly the case to-day. A disproportion is noticeable between the large sum devoted to the space segment and that allocated to the assessment and use of the data for cooperation.

Will Europe, both Europe, that of the "Unique market" and that of our "Common Europe home" which are being built, be able to act as a coordinating agency and facilitate an extended use of data and products emerging from an extended cooperation ? Will Europe bring new tasks to public and private agencies in each country and create new public facilities contributing to the general economy and management of a modern country, considered as an integral part of a wider european community ? Will Europe be able to share this wealth of data with developing countries, which are moreover poorly mapped countries claiming such products ? Or, on the reverse, will ECC set up quotas for data from space for each country and for each national earth observation system, in a way similar to the malthusian quotas for milk, in order to save money at short notice, like in a new data-applied version of the past witch hunt policy ? Europe has a formidable task ahead of her. Who will decide whether there is too much data or not ? Who will replace some possibly missing data the day when they would be required for a better understanding of our planet ? We are condemned to be more and more data-dependent, but it does not seem that we are really overwhelmed with data from space sensors, in comparison with the volume of other land related information which are collected by other means, despite sometimes, their lack of efficiency, nor even in comparison with the volume of data collected from aerial surveys.

REFERENCES

[1] BRANDENBERGER, A. : Status of the world topographic and cadastral mapping. Proc. Int. Forum on Instrumentation and Geographic Information (FI3G, Lyon (F), Vol. 1, pp. 11-25 (CNIG, 136 bis Rue de Grenelle, 75007 Paris (F)).

[2] DUCHER, G. and PLU, D. : Availability of Aerial Photography and Space Images ; the Photothèque at the I.G.N. (F). Photogrammetria (PRS), 43 (1988) pp.83-100.

[3] DUCHER, G. : The current involvement of I.G.N. in photogrammetry. The Photogrammetric Record, 13 (73), pp. 111-126 (April 1989).

[4] DENIS, P. and GUILLOUET, G. : L'action de Développement, Evaluation, Formation sur SPOT (ADEF) : évaluation de la restitution photogrammétrique. Bulletin d'Information de l'I.G.N. n° 57 (1989), to be published.

[5] LANDSAT data users notes n° 35 - March 1986, pp. 1-2.

[6] SPOT News-letters N° 11, pp. 9-12, March 1989.

[7] MACDONALD, Alastair. : Is there any basis for BASIS ? Survey and Mapping 89. Univ. of Warwick (G-B) 17-20 April 1989.

[8] BURNSIDE, C.D. : Report on ISPRS Commission I (Kyoto Congress). The Photogrammetric Record, 13 (73), p. 9 (April 1989).