

# **EAIR: Special Interest Group (SIG) on Exploiting Data Repositories**

Marcel Herbst & Mantz Yorke

December 1, 2006

## **Contents**

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>The Rome Forum Session</b>	<b>2</b>
<b>3</b>	<b>Outlook 2006–07</b>	<b>4</b>
3.1	Core Group in Support of the SIG . . . . .	5
3.2	Publications . . . . .	6
<b>4</b>	<b>EAIR-Forum in Innsbruck (2007)</b>	<b>7</b>
<b>5</b>	<b>Focus on Content</b>	<b>7</b>
5.1	Existing Data Repositories . . . . .	8
5.2	Problem Areas and Associated Methodologies . . . . .	8
<b>A</b>	<b>The Gestation Process</b>	<b>9</b>
<b>B</b>	<b>Sources on Institutional Research</b>	<b>13</b>

## **1 Introduction**

We had a good session of our constitutional meeting in Rome, but we had a somewhat deficient follow-up. September 18, 2006, Marcel had issued a

call to “jot down your thoughts and wishes regarding our SIG after you had returned home”. A few of you responded. Subsequently, a new note was issued. However, both Mantz and Marcel were busy, apparently, and it is only now that we can send out this note.

In the meantime you will also have received the call for paper proposals for the EAIR 2007 Forum to be held in Innsbruck<sup>1</sup>. The deadline for submitting paper proposals is January 9, 2007. The Innsbruck Forum has seven tracks, all of which would welcome papers that have an empirical bent or would exploit data repositories. Perhaps two tracks are particularly oriented toward empirical analysis: “Constructing Meaning from Performance Measures” (Track 5), and “Matching Staff, Structures, and Resources” (Track 7). If you happen to submit a paper proposal with an orientation that would fit our SIG, please let us know (we might want to coordinate the matter). We shall plan a special slot for our SIG at the Innsbruck Forum.

This note recapitulates our session at the Rome Forum (see Section 2), provides an outlook on our work during the academic year 2006–07 (Section 3) and regarding the upcoming Innsbruck Forum (Section 4), summarizes possible foci of concern (Section 5), and adds two appendices: one that describes the gestation process of our SIG (Appendix A), and the other that refers to two homepages which refer to data sources (Appendix B).

## 2 The Rome Forum Session

Mantz Yorke, Victor (Vic) Borden and David Kane summarized the findings of our SIG-Session<sup>2</sup>. On the part of Mantz, it is a very ‘top-level’ approach since Mantz was not privy (nor Marcel) to much of the discussion within the groups. On the part of Vic and David, only one group was the focus. Two main (not necessarily disjointed) themes appeared to stand in the foreground:

**Theme 1:** benchmarking and comparisons (at levels ranging from within-institution to international);

**Theme 2:** information for management purposes (management information tools).

Within these two themes there were a number of issues, including:

---

<sup>1</sup><http://www.eair.nl/innsbruck/>

<sup>2</sup>eMails of September 10 and 19, and October 6, 2006, respectively.

- What data are actually available? How are data defined and structured?
- How can data (from various sources) be mapped across from one context to another — and produce meaningful comparisons? How can they be used for longitudinal as well as cross-sectional analyses?
- How long do data remain valid? How fast is the rate of change in higher education? How necessary are historical records?
- What are the obligations of the ‘owner’ of the data vis-à-vis the public and the potential ‘user’? How can data access be insured? What ethical issues pose themselves?

There are various angles on the two main themes (above). There are issues of policy at the level of institutions, and implications for quality assurance and enhancement. Considerations at the level above the individual institution include state policy and marketisation of higher education. There is also a need to be aware of “what’s not there” in databases. And there is perhaps a need for creative approaches to the use of databases.

At our session, we had three groups discussing (overlapping) issues relating to our SIG. David Kane observed a “diversity of interest among those present”. Presumably, this applies to all three groups. A summary of such issues might look as follows:

**Group 1 (Vic Borden):** What data are available where: an inventory of data systems. Data definitions and standards: a view on the consistency and reliability of the contents of extant data systems. What are we or should we be trying to measure? — a view on the conceptual underpinnings of our attempts to use data to measure higher education structures, operations, and outcomes. What ethical issues arise from the use of data repositories? Who pays for data collection (and do the outcomes of this collection influence funding)? Data use in risk management and for purposes such as the research assessment exercises (with the use of metrics rising up the agenda). Tracking of costs, within institutional finance systems. How can we share? — a technological approach to improving the availability and accessibility of extant data and benchmarks. How do we use the data effectively? — a methodological approach to effective and appropriate analyses of extant data.

**Group 2 (Rosalind Pritschar):** Benchmarking as a contribution to assuring the quality in higher education (HE), but having political overtones. A potential conflict exists between enhancing the quality of HE and some political objectives (such as: should the 'best' be rewarded, or should the policy aim be to bring weaker performers closer to the standard of the 'best?'). How does one define quality standards (by levels of performance, by degrees of performance improvements, etc.)? Does one distinguish between established performers (the static view) and potential performers (the dynamic view)? Problems exist with data quality in some countries, perhaps stemming from criteria that were inadequate. Who is collecting (i.e. 'owning') the data, and for what purposes? Are data made public? How is trust established between the data collection agency (the final 'owner' of the data) and the 'observed' institution?

**Group 3 (David Kane):** Questions arise regarding what it is that one is trying to measure, and whether some things are intrinsically measurable in a strict sense. The need to ensure that "like for like" measurement is achieved, otherwise benchmarking and comparisons are rendered suspect. Consensus might be possible (but it would need some work to achieve this). There is also the problem of data accumulations (e.g. to compress [or transform] vectors of data into scalars [i.e. indicators, scores, ranks]). Issues where comparisons would be valuable would include (i) access to HE, (ii) learning, (iii) other benefits arising from HE.

### 3 Outlook 2006–07

At the Rome EAIR-Forum, we decided that in order for the SIG to become operative, we ought to pay attention to the following:

- forming of a Core Group in support of the SIG;
- publications (within TEAM, within some new publication series of EAIR, etc.);
- activities planned for Innsbruck (2007).

### 3.1 Core Group in Support of the SIG

A core group (CG) was formed at the Rome Forum. The formation of this CG was informal, that is, people who think they are in a position to contribute and would like to join should be able to join anytime (no formal procedure is required). In a few years, this informal procedure might be replaced by a more formal procedure (if necessary)<sup>3</sup>. According to our notes (which may not be trusted), the following persons have agreed to serve in the CG:

- Borden, Victor (US), Indiana University (vborden@indiana.edu);
- Delaney, Anne Marie (US), Babson College (delaneya@babson.edu);
- Hoekstra, Peter (NL), Universiteit van Amsterdam (j.p.hoekstra@uva.nl);
- Herbst, Marcel (CH), 4mation (herbst@4mat.ch);
- Hugentobler, Urs (CH), ETH Zürich (urs.hugentobler@fc.ethz.ch);
- Kane, David (UK), University of Central England in Birmingham (david.kane@uce.ac.uk);
- Longden, Bernard (UK), Liverpool Hope University (longdeb@hope.ac.uk);
- Pritchard, Rosalind (UK), University of Ulster (r.pritchard@ulster.ac.uk);
- Vermeulen, Pieter Jacobus (SA), University of Pretoria (pieter.vermeulen@up.ac.za);
- Yorke, Mantz (UK), Lancaster University (mantzyorke@mantzyorke.plus.com).

The CG of our SIG is composed of people having differing interests, and they may assume different tasks in the future. In forming the CG we have, initially, the following tasks in mind:

---

<sup>3</sup>Many small organizations or publication boards operate successfully with informal setups: their members know each other (after they have met at meetings or have exchanged paper drafts via eMail).

- **Steering or Coordinating Group (SCG):** For the time being, Mantz Yorke and Marcel Herbst assume this function. Other people are welcome to join (perhaps, one or two additional members are needed).
- **Area Representatives (AR):** CG-members and members of our SIG represent different geographic regions (countries), they are knowledgeable (to some extent) regarding the data repositories on — and data usage within — their home countries, and they can act a AR. Currently, the CG has AR of the following nations: CH, NL, SA, UK and US. Members of our SIG may in a position to represent countries not represented by members of the CG.
- **Thematic Issues (TI):** Members of EAIR focus on themes or issues they are involved with. Normally, issues of governance, management, evaluation, quality, formation, etc., are addressed within a Forum, but TEAM or its Editorial Board, for instance, does not have a specific thematic orientation. The SIG, and the SCG, should adopt the same position: members of the CG will just have their individual interests and are in a position to represent those they have some experience with.
- **Methodology Group (MG):** As stated earlier (see the note of August 21, 2006), a certain deficit regarding methodology has been decried. Hence, the SIG should attract and involve people with a methodological bent and a desire to fight against that deficit. Marcel shall act (for the time being) as the coordinator of the MG. Please contact Marcel if you would like to join this group.

The members of the CG should communicate among another, and the more formal notes of communication should be aided by the SCG.

### 3.2 Publications

Presumably, all new SIG's will have an impact on the publication pattern of EAIR. Currently, TEAM pretty much publishes papers presented at the annual forum (with exceptions). As of now, TEAM does not have declared thematic issues (although the editors try to achieve some thematic coherence). In the future, perhaps, TEAM might try to bundle papers under one overall theme. EAIR also discusses a new publication series which could better serve a thematic orientation and which could focus on the interests of the individual SIG's.

Eventually, our SIG should foresee to produce publications. We do not know when this will be the case, but we hope we can start envisioning such activities after the EAIR-Forum in Innsbruck (2007).

#### 4 EAIR-Forum in Innsbruck (2007)

The current year ought to be used to consolidate our SIG: to exchange ideas and problems among its members, to recruit new members, to help members in their daily work, and to prepare the activities for the EAIR-Forum in Innsbruck:

- A special session of our SIG shall take place at the EAIR Forum in Innsbruck (analogous to that of Rome).
- We plan a number of 10-minute slots where members of our SIG present Case Studies regarding their experiences with — or use of — data repositories. Short papers on some of these Case Studies shall be included in future publications of our SIG. Please let us know until <Date> if you would like to make such a presentation.
- Following a suggestion by Vic Borden, CG-members and SIG-participants are invited to collate information on data repositories they consider important (an international data system inventory; for a possible format of this collation, see 5.1). This should provide us with “annotated URL’s” for sites related to extant higher education data systems. This information would be made available in a online publication (a resource page) that can be offered to EAIR members through the web site. Keeping this information updated will be another problem. We would like to make a first version of this information available prior to the Innsbruck Forum, if possible.

#### 5 Focus on Content

All members of the CG of the SIG, and in particular AR and members of the MG, will have to focus on existing data repositories, problem areas (or issues) and associated methodologies. Members of the SIG will have to deposit their interest or concern.

## 5.1 Existing Data Repositories

AR might focus on data repositories within their own country or jurisdiction and report according to a scheme as follows:

- authoring agency, location, web address;
- publications (in paper) which report these data;
- design audience, and audience which might profit from using these data;
- access (restricted or not), costs of accessing information or downloading files;
- how data are stored (format, etc.), query language;
- major publications and research reports which use these data.

In due time, we can report on data repositories and their use. This might help ease to find access by those not directly (professionally) involved.

## 5.2 Problem Areas and Associated Methodologies

Earlier, Marcel stated that “issues drive data collection and exploration” (see Note of August 21, 2006, page 3). Furthermore, data exploration is closely linked to methodological questions. EAIR-members are interested in or know about problem areas: they are confronted with issues or problems that need to be addressed or solved. But many sensible data explorations or analyses are not made if we lack the methodological tools. And conversely, if we do not see methodologies applied in a constructive way, we lack the incentive to become methodologically interested.

Data contained in higher education data repositories can frequently be seen organized (implicitly) along the following categories:

- input-indicators (student achievement as measured for incoming students, socio-economic background of students);
- output-indicators (such as research publications, graduation or dropout rates, student achievement as measured for graduating students);
- outcome-indicators (such as citations, first destination of graduates);
- throughput-indicators:

- student-faculty ratios, staff-faculty ratios, time-allocation of faculty (teaching, research, services, etc.), faculty distribution (by rank);
- part-time versus full-time studies of students, geographic mobility (foreign students, Erasmus, etc.) and disciplinary mobility of students;
- employment and promotion patterns (age when employed or promoted, age distributions, replacement rates, etc.); library holdings).

Some data are difficult to place in this scheme, because it would depend on the viewpoint. Research grants generated could be seen as an output-indicator (i.e. an output of the past research activity or the indicator of research reputation) or as an input-indicator (i.e. an input for future research). Student achievement measures are input- or output-indicators depending on whether students enter or leave an institution.

Data found in repositories are placed there because they are seen as part of an important problem situation or issue. People exploiting data repositories may not have the same problem situation or issues in mind when trying to extract data for research purposes. Hence, we have three issues:

- exploiting data repositories to analyze issues which stood in the foreground when designing the repository (this is the ordinary institutional application);
- exploiting data repositories to analyze issues which might not have stood in the foreground when founding the repository (this is the extraordinary application);
- trying to address issues for which data are (partially) lacking.

## A The Gestation Process

A paper by Mantz Yorke kindled the thought to found a Special Interest Group (SIG) on Exploiting Data Repositories by Marcel Herbst. For the record, the following four eMails document the early thoughts in this respect.

**Review of Riga Paper of Mantz Yorke (early January, 2006):** Mantz' Riga paper was sent by Barbara Kehm (an Editor of TEAM) to Marcel (a member of the Editorial Board of TEAM)<sup>4</sup>. In his review of Mantz' paper Marcel stated the following :

The paper addresses three problems (according to the author), but I shall refer to two: (i) there are data available (in publicly accessible data repositories) ready to be analysed; and (ii) four examples of possible data and analyses are given. I think the paper is disjointed. Both themes — i.e. (i) and (ii) — are important, very important in fact, but they should not be presented together (in a short paper).

I would devote one full paper to the first theme, would list data repositories which are under-used or under-researched, would list possible problem areas or issues which could be addressed, or perhaps even possible analyses (with a focus on methodology). It has been my observation, over the years, that higher education or institutional researchers are rather weak methodologically (regarding statistics or operational research), and such a survey could invite practitioners or academics to explore new avenues. In fact, I myself considered writing such a paper (or even a book), but because I did not find any co-authors, I gave up on the idea (for the time being).

The examples the author gives should not serve as examples. The cases should stand for particular analyses or methodologies — or for problems which one might encounter in doing survey work or analysis, etc. In other words, I would welcome here another focus: again, a new paper.

I would welcome the author to redirect his attention (and to write two new papers). I can readily understand, however, if he would want to ignore my advise. The paper cannot be rejected because it is bad (but the two new papers could be really enlightening, and they would not depend on each other)<sup>5</sup>.

**Answer of Mantz Yorke (early January, 2006):** Mantz responded to Barbara Kehm and she related Mantz's position statement to Marcel:

Thanks for the comments [by Marcel]. I think that the reviewer makes some valuable points about what might be written on the subject but, having re-read the piece, I don't agree that it is disjointed. It does what it set out to do, within the confines of a Forum paper. Granted, with twice the space, I might have made more of the points addressed by the reviewer: my aim however was simply to illustrate the potential that is 'lying around', un/underexploited, and by doing so to encourage others to look around their particular 'patches' for IR opportunities.

<sup>4</sup>All papers of an EAIR-Forum to be included in TEAM are reviewed, and this applies also to papers by editors or members of the Editorial Board (such as Mantz Yorke).

<sup>5</sup>The paper was subsequently published. See: Mantz Yorke, "Gold in them There Hills? Extracting and Using Data from Existing Sources", *Tertiary Education and Management*, Vol. 12, No. 3, September 2006, pp. 201–213.

The point about IR personnel's expertise with statistical data is pretty valid for the UK, where research methodologies have latterly tended towards the qualitative, and may apply elsewhere in Europe. I don't think it applies in the US, where IR people routinely crunch numbers (though the problem there may be that the crunching is done uncritically — I recall one published statistical paper in which the proportion of the variance explained by the measures used was a glorious 19%!).

I think that the reviewer makes more of the three propositions than I intended — indeed to couch them as 'problems' implies an examination of the propositions along the lines of his/her comments. My 'take' on the propositions was that these are things I have found through practical experience, and for which I can provide some evidence. Hence my assertion of a lack of disjointedness, in that propositions 1 and 2 [which Mantz refers to in his paper] are linked through the examples. If that link withstands scrutiny, then proposition 3 follows.

The difference between me and the reviewer is one of perspective. I suggest that, from our different perspectives, we are both 'right'. I'd quite like to take up the challenge thrown up by the reviewer, but could only do this when I've cleared a pile of other things from my desk (and that will take a longer time than you'll be wanting for TEAM). Maybe the answer is for the reviewer and me to get together . . .

**Reponse by Marcel (January 18, 2006):** Marcel contacted Mantz (and Barbara) with the following:

Dear Mantz (and dear Barbara):

Yes, I had to review your paper. I was in a kind of a dilemma. I thought the two themes which I extracted to be very important, and I thought your paper doesn't properly address those (you had a different aim in mind). I presume you will revise your paper anyhow (because of the other reviewer's comments). Here, I make you (and Barbara) now a new proposal.

As I already indicated in my review of your paper, I am not that happy with the papers submitted to TEAM (and implicitly with the research conducted). There are very few papers which have an empirical or quantitative bent. Survey analyses are the most common of quantitative papers, but there are practically no econometric or statistical analyses focusing on the many data repositories which exist in a range of countries (this applies also for the institutional researchers in the US, not for economists, however). For this reason, I welcomed your observation that data repositories are underused.

The second aspect is methodology. Here, people may have limited ideas concerning what can be extracted from data. So, examples might help to encourage people to cooperate with others who are perhaps a bit more familiar with quantitative analyses.

I propose to schedule a special session at the EAIR Forum in Rome to discuss these two issues. The aim should be to form a group of people who would take on the responsibility for an entire issue of TEAM devoted to these matters. Since the deadline for submitting papers etc. [to the Rome-Forum] is

past, we would have to contact the Forum organizers for help (we would need a double-session). We could also place a call to attract other presenters. [...] I think we should do this only if TEAM agrees (in principle). TEAM normally contains 5 papers (hence, we would need roughly 5 authors). I could present in Rome a survey of methodology (and you as well). I noticed that the UK has very fine data repositories (which I would like to analyse), as has the US. The situation is much more difficult in other countries (access, language, etc.). Here, we should have people familiar with the situation in Germany, the Netherlands, Scandinavian countries, France, etc.

What do you (and Barbara) think? I do not want to push you; there is no rush. But we shouldn't forego the opportunity to discuss this in Rome (if we would like to pursue this road). I shall be off on vacation from January 21 to February 3 (I may be able to read eMails).

With kind regards,

Marcel

**Reponse by Mantz (January 18, 2006):** Mantz answered Marcel with the following:

Dear Marcel:

Thank you for your message and for your earlier anonymous comments. What you say chimes with some of my concerns — some of which I was unsuccessful in pressing when on the EAIR Executive, such as exploiting the international membership of EAIR to do things on at least a cross-European basis. EAIR doesn't do this (well, it does to a limited extent, since the international seminars do bring people together to discuss common issues) — my concern has been about bringing people together to work together on issues that are of broad interest but which require colleagues in different countries to contribute positively. In the context of the present discussion, this might be analysis of data relating to, say, student completion, student satisfaction and so on (you can see my kinds of interest here, and would no doubt want to put other things (e.g. econometric?) into this pot).

Stage 1 might be to list a series of problems or issues that are amenable to statistical analysis using databases that already exist (there would be a need to identify them, which is why knowledgeable people from different countries would be needed. So I suppose that this would suggest some sort of international brainstorming event which might form the heart of a Forum session. Maybe something less formalized than what I understand you to be wanting to do.

[...]

Then having looked at what might be researched, and what data resources were available, Stage 2 would be to do some comparative work on the problems/issues that strike a chord with Euro-colleagues and with policy-makers.

Notice that, for the time being, I have focused on Europe rather than the wider world (which your remarks cover). Not that I'm thinking 'fortress Eu-

rope’ — more that there might be some funding to do something if it has a strong Euro focus. That isn’t to say that it wouldn’t be valuable, for the purposes of comparison, to bring — say Australia and the US — into the frame as well.

With best wishes for 2006,

Mantz

## B Sources on Institutional Research

We have not yet attempted to collate a bibliography — or online resources — on institutional research (IR). For the time being, the website of our parent organization AIR is a good way to start:

- [www.airweb.org/](http://www.airweb.org/)

It contains numerous links to web-based IR resources. Naturally, the collection is heavily focused on the US, and it would be our turn to develop a corresponding listing of non-US — i.e. European and outer-European — resources.

Mantz refers to the *Primer for Institutional Research*, issued by AIR and edited by William E. Knight (2003)<sup>6</sup>. It contains, among others, a valuable chapter on “Using National Datasets for Postsecondary Education Research” by John H. Milam which you can find under Milam’s homepage:

- <http://highered.org/>.

---

<sup>6</sup>See: [www.airweb.org/p.asp?page=429](http://www.airweb.org/p.asp?page=429).