Creation of a BI environment from scratch with open source software, a practical case

Thierry Dautcourt

To cite this version:

Thierry Dautcourt. Creation of a BI environment from scratch with open source software, a practical case. EUNIS - European University Information Systems Congress, Umeå universitet, Jun 2014, Umeå, Sweden. hal-01009396

HAL Id: hal-01009396
https://hal.inria.fr/hal-01009396

Submitted on 23 Jun 2014

Creation of a BI environment from scratch with open source software.

A practical case – Thierry Dautcourt[1]

[1] Inria, 54600 Villers-lès-Nancy, Thierry.Dautcourt@inria.fr

This presentation examines a practical case and the methods used to build a BI (Business Intelligence) environment, in a Research's Institute lacking initial BI maturity. We describe here some conditions for success and the choice of open source software to build the technical BI environment.

Context

The French Institute for Research in Computer Science and Automation (INRIA) is a public research body fully dedicated to computational sciences. It is organized into 210 research teams split across eight Inria centres. In 2010, our management requested our own BI environment. The main purpose of the BI environment was to produce « automatically » the major performance indicators related to our quadrennial contract signed with the Government (Ministry). At the time, these major indicators were produced by aggregating heterogeneous spreadsheet files and with a lot of manual work for collecting, aggregating and reconciling the data.

How to proceed ? Where to begin ? The field of contract indicators covers a very large domain from human resources, and financial to scientific production. In addition, the quality of data was uneven and the culture of data management rather low. The approach was to work on three axes: identifying one strategic indicator, finding people to work with and building a technical environment for people to access analytic data.

Focus on one strategic indicator

The first axis was to focus on one major indicator among many. We choose one whose values were not computed. This indicator is the rate of deposit[1] in the open access system of scientific publications produced by researchers. As Inria is strongly involved in the «Berlin Declaration[2] » the goal was to provide full open access to knowledge in the Sciences and Humanities. For this measure, we relied on two sources of data: the publications present in the open archive system (HAL[3]), and the existing publications in the team's annual report activity[4].

Finding the people to work with

The second axis was to find « open minded people » to work with. The subject of publications was an opportunity, as people working in scientific and technical Information are usually working in a network and are very involved in data quality and data cleaning processes.

Create a BI environment with open sources

The third axis is the technological environment. We didn't have a budget for software licences. Instead we used three open source software packages, *Talend Data Integration* for the ETL (Extract, Transform and Load) and *SpagoBI* for the BI web server. Data are stored in MySQL databases.

Talend Data Integration[5] (community version) was used for building the data warehouse. This tool is very accurate for incremental conception. The different phases of transformation are mostly done without any lines of programming but with connecting high-level components in sequence. However, this does not eliminate method. Evolution over time is easy to manage and costs of modification are small (1 or 2 days), depending on the size of modification.

[1] Here, the rate of deposit of the publication references, another measure is the "full text" deposit.

[2] Signed by Gilles Kahn in 2004, openaccess.mpg.de/*Berlin-Declaration*

[3] hal.inria.fr

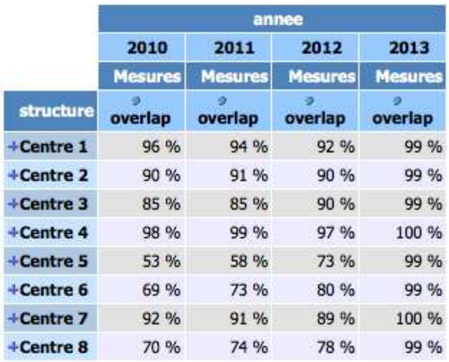
[4] *raweb.inria.fr/rapportsactivite/RA2012/index.html*

[5] *www.talendforge.org*

The ETL is used for extracting, homogenising, denormalizing, and matching data in a data warehouse. We must also manage errors issued by harmonizing data of two sources, an open archive and an annual report. The goal was to build a *multidimensional dataset* (cube) on subject “publications”.

SpagoBI[6] is the second tool for providing a web access to data (and errors for correction), for exploring and consulting analytical views. This is a full open source Business Intelligence platform that uses many tools as analytical engines. Our first major interest was about OLAP (On-Line Analytical Processing) cube.

The dimensions used for analysis are the organizational dimension (team project and research centre), time dimension (year) and other dimensions specific to the subject area, such as the typology of publications dimension. The measure, rate of deposit, is computed accounting for the overlap between publications in the team’s annual activity report and publications in open access archive HAL. *Figure 1* presents the measure by two dimensions: years and research organisation, here INRIAs 8 research centers.



	annee			
	2010	2011	2012	2013
	Mesures	Mesures	Mesures	Mesures
structure	overlap	overlap	overlap	overlap
+Centre 1	96 %	94 %	92 %	99 %
+Centre 2	90 %	91 %	90 %	99 %
+Centre 3	85 %	85 %	90 %	99 %
+Centre 4	98 %	99 %	97 %	100 %
+Centre 5	53 %	58 %	73 %	99 %
+Centre 6	69 %	73 %	80 %	99 %
+Centre 7	92 %	91 %	89 %	100 %
+Centre 8	70 %	74 %	78 %	99 %

Figure 1 View of publications references deposit's rate

The others conditions to succeed (and not least!)

The platform allows for monitoring the deposit of publications over time, in open archive. But this is not enough. Additional effort focused on two directions: 1) Services to researchers for providing a dynamic html frame on publications ready to integrate in team or personnel web pages[7]; 2) An accompanying program from scientific and technical information staff, helping researchers to deposit publication in the HAL archive. The indicator helped highlight “low deposit teams”, and to uncover the reasons for their poor commitment.

For many years, to support research teams in the production of their annual activity report, a document skeleton is produced using data from our IS (information system). For the 2013 annual activity report, for the first time, all publications were automatically fed in the team annual report skeleton.

This is also the case for other data. Each year, more data are added automatically from the data warehouse in the “activity report skeleton” such as *team participants, funding’s, International, European and National relationships with Industry and University or Research Centre*.

In 2014, our BI environment integrates about 12 analytics cubes and 40 views. Increasingly, the data warehouse became a key part of our IS. The business approach and agile construction of data warehouse help the dialog between business requestors and IT people for building IS. The exposition and reuse of data by a larger population increases the robustness of the Institute’s data and along the way, quality indicators.

References

Kimball, R., & Ross, M. (2002). *The data warehouse toolkit: The complete guide to dimensional modeling (2nd ed.)*. New York, NY, USA: John Wiley & Sons, Inc.

Whitehorn, M., Zare, R., & Pasumansky, M. (2006). *Fast track to mdx (2. ed.)*. Springer.

Author fr.linkedin.com/pub/thierry-dautcourt/52/69a/11

[6] *spagobi.org*

[7] Example: *www.inria.fr/en/teams/alice/%28section%29/publications*