Towards Automated Enterprise Architecture Documentation: Data Quality Aspects of SAP PI

Sebastian Grunow, Florian Matthes, and Sascha Roth

Abstract Well executed, Enterprise Architecture (EA) management is commonly perceived as a strategic advantage. EA management sermonizes IT savvy firms to take profound decisions based on mature EA information. As of today, gathering required information, i.e. documenting the EA, is regarded both, time consuming and error-prone. As a reaction, recent approaches seek to automate EA documentation by extracting information out of productive system environments. In our recent work we showed that a particular Enterprise Service Bus (ESB) namely SAP Process Integration can be used to extract EA relevant information. As a next step towards automated EA documentation, this paper analyzes the quality of the data stored in SAP Process Integration systems in practice. Survey results of 19 industry partners on 4 continents are presented.

Key words: Enterprise Service Bus (ESB), SAP PI, data quality, automated Enterprise Architecture documentation

1 Introduction and Motivation

Enterprise Architecture (EA) management is commonly perceived as strategic advantage [16]. Approaches from academia and practitioners, e.g. [23, 3], teach to take profound EA related decisions based on mature EA information. These approaches commonly start an EA endeavor by capturing the current state (as-is) of the EA and create stakeholder-specific visualizations for analyses [17, 10]. As of today documenting the EA requires manual collection of data and thus is regarded as an error prone, expensive, and time consuming task. As a reaction, researchers and prac-

Software Engineering of Business Information Systems (sebis),
Technische Universität München, Garching b. München 85748, Germany
{grunow,matthes,sascha.roth}@in.tum.de

titioners [5, 7] seek to automate EA documentation. These approaches focus on extracting information out of productive system environments, but their data quality aspects are not addressed by the authors.

In our recent work [4, 9] we showed that a particular Enterprise Service Bus (ESB) implementation can be used to extract EA relevant information. We investigated an ESB since it can be "considered as the nervous system of an enterprise interconnecting business applications and processes as an information source" [4]. In our analysis we compared concepts contained in the ESB (e.g., interface descriptions, participatinng applications and systems) and the focus was put on the evaluation of the coverage degree to which data of a productive system can be used for EA documentation, i.e. we focused on a model mapping rather than data quality aspects. Results published assume the best data quality (complete, correct, up-to-date) within the productive systems, i.e. data quality aspects are neglected entirely. When applying the idea of an automated EA documentation (cf. [9]) to productive system environments, the actual data quality has a high influence on the outcome of such an endeavor.

In this paper, we analyze data quality aspects of a particular ESB system, namely SAP PI. Analyzed data quality aspects indicate whether or not those systems can be used for an automated EA documentation in practice. To support our recent research question, i.e. 'to which extent can an SAP PI system be used for an automated EA documentation?' (cf. [4]), we conducted a survey among 19 industry partners distributed on 4 continents. These results are a next step towards the practical application of an automated EA documentation.

The remainder of the article is structured as follows: Section 2 presents related work followed by Section 3 that reports results of an EA data quality assessment of ESB systems in productive environments. An interpretation of these results with respect to our research endeavor 'automated EA documentation' is given in Section 4. The paper concludes with an outlook in Section 5 and outlines further research directions.

2 Related Work

Existing EA frameworks covering inter alia *The Open Group Architecture Framework* (TOGAF) [20], *The Integrated Architecture Framework* (IAF) [22], and *Enterprise Architecture Planning* (EAP) [18] commonly do not detail how to acquire and incorporate EA knowledge. Only few approaches considering the documentation of the status quo could be identified. However, the descriptions usually take place on a high abstraction level without consideration of concrete process tasks. For instance, TOGAF suggests the usage of existing architecture definitions as a starting point, which if necessary, can be updated and verified. In case such information is unavailable the collection of data *"whatever format comes to hand"* [20] is advised by TOGAF.

Moser et al. [13] give a first idea for an automated tool-aided collection process by introducing a set of EA process patterns, one of which is called *Automatic Data Acquisition / Maintenance*. The authors propose a process aimed at automatically collecting data from various sources converted into an EA information model instance. However, the considerations do not detail possible information sources including data quality thereof.

Based on identified requirements on an automated documentation process, Farwick et al. [6] develop a basic structure of an automated maintenance process comprising the collection of data as well as the propagation of changes. However, when it comes to implementation details the authors refer to future work.

A more detailed look on an implementation is taken by Buschle et al. in [5], whereby NeXpose, a vulnerability scanner aimed at determining weaknesses within the system landscapes is used. Apart from weaknesses the scanner also collects information about the underlying systems landscape which subsequently is mapped to an EA information model. Hence, information on existing services, installed software, and used operating systems could be gathered. While the information coverage, i.e. the extent to which the demanded EA information can be determined, is considered within the publication the usefulness of the collected data is not discussed leaving open questions – as to the correctness and completeness of the data, for instance. Instead of using a vulnerability scanner, [4] employ an ESB. While the degree of coverage to which data of a productive system can can be used for EA documentation is thoroughly analyzed data quality thereof is not evaluated.

To the authors best knowledge there is no existing research analyzing data quality aspects of ESB systems and in particular SAP PI systems in productive IT environments.

3 SAP PI Data Quality Assessment: Evaluation of the Survey Results

In line with Steen et al. [19] we observed the EAs to be an important starting point for analysis, design, and decision processes. For this to work, EA information must provide an *accurate*, *correct*, and *up-to-date* model of the real world [6]. Accordingly, aiming at developing an automated EA documentation process the suitability of an ESB system as an information source is not only determined by information content but also by the quality of the data saved in the system, e.g., Are the data up-to-date? Are the attribute values correct? and Are the data consistent?.

Subsection 3.1 gives an overview of SAP PI. In order to gain a deeper insight into the data quality of a SAP PI system in practice we conducted interviews with SAP PI experts and an online survey with EA practitioners. Overall, 19 EA practitioners participated in the subsequent online survey and could be identified as responsible for an SAP PI system. Thereby, questions about the data content of SAP PI were assessed in terms of selected quality dimensions (see Subsection 3.2). In Subsection 3.3 we present the survey results in greater detail.

3.1 Overview of the SAP PI data

SAP PI is not a single module but rather a conglomerate of various components, which are not independent but stand in relationship to each other (see Fig. 1). The *System Landscape Directory* represents a central provider of landscape information comprising information about installed and installable software as well as technical details about the underlying infrastructure. Designing, creating and maintaining the interactions between the applications takes place in the *Enterprise Service Builder* which includes amongst others interface descriptions, messages, and exchanged data types. In the *Integration Builder* configurations of communication relationships at run time map *Enterprise Service Builder* elements to the actual execution environment. In order to test and monitor an SAP PI installation the *Runtime Workbench* offers a central entry point putting in place various tools. Finally, the *Integration Server* is responsible for processing incoming messages from sending applications, applying routing and mapping rules, and finally forwarding them to target systems.

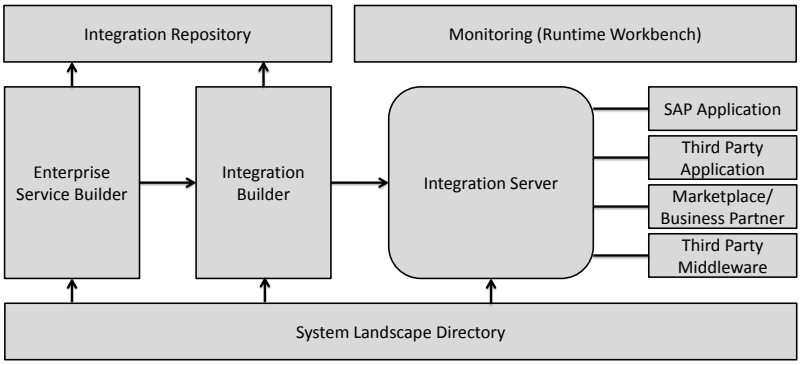


Fig. 1 Architecture of SAP NetWeaver Process Integration [14]

3.2 Data Quality Dimensions

An examination of literature reveals a high variety of quality definitions [11, 1, 8, 12]. For instance, Bednar et al. [15] make a distinction between four views on quality: *quality as excellence*, *quality as value*, *quality as conformance to guidelines* and *quality as meeting or exceeding customer expectations*. The first two views may turn out to be problematic as the assessment of excellence involves a high degree of subjectivity and the determination of a value is highly influenced by a monetary perspective while neglecting further criteria [11]. In contrast, ISO 9001:2008 [2] specifies quality as the "totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs" [1]. Quality respectively means

the fulfillment of required characteristics, also referred to as quality attributes. To establish a link to the first definition both views *quality as conformance to guidelines* and *quality as meeting or exceeding expectations* can be dissembled into a set of quality attributes to be fulfilled, e.g., completeness, accuracy and correctness.

Depending on the research area different frameworks proposing various quality attributes have been developed in an attempt to assess quality, such as software quality ([8, 12]), data quality ([21]), information quality ([11]) and even Enterprise Architecture quality ([6]). Consequently, the first challenge to overcome is to reduce the broad range of existing quality attributes to the essential amount to gain reasonable coverage of the term quality. As the focus lies on the inserted data in SAP PI rather than on the way they are saved, quality aspects regarding the underlying data model as well as its usability, e.g., simplicity, relevance of the data, perception from the user's perspective, semantics, etc. are considered as given and neglected.

A comparison of the different taxonomies reveals many terms including completeness, correctness, and actuality are common to most of them indicating a consensus in research and industry. In addition, attempting among others to determine Enterprise Architecture quality dimensions by conducting a survey Farwick et al. [6] identifies completeness, correctness, and actuality to be ranked highest. Accordingly, to make qualitative statements about the generated EA information the analysis of the SAP PI data quality in terms of these quality dimensions provides a good starting point. A list of the quality dimensions chosen can be found in Table 3.2.

Quality Dimension	**Description**
Completeness	The extent to which the expected data are provided according to the SAP PI specification
Correctness	The degree to which the SAP PI data reflect the real world and fulfill the SAP PI guidelines
Actuality	The degree to which the data are up to date

Table 1 Considered data quality dimensions [11, 1, 8, 12]

3.3 Assessment of SAP PI

As no research concerned with the quality of ESB systems in general and SAP PI systems in particular could be identified, we conducted an online survey aimed at evaluating the quality of data contained in SAP PI systems in practice. The survey was opened within 45 days. 45 SAP PI experts started the survey, 19 fully completed, 24 partly completed it whereas 2 only answered the first two questions. The last two respondents are neglected in the subsequent analysis.

General information about the respondents: 50 percent of the respondents reside in Asia, 30 percent in Europe and the remainders are distributed equally over North

and South America. Out of all respondents 67 percent claimed to work for a consulting company. This circumstance favours the results of the survey as the answers include the experience about the situation in more than one organization.

Perception of the overall quality: All respondents rate their data quality as nearly perfect (80 percent) or even perfect (20 percent) which is partly attributable to high correlation between quality and functioning of the system. Incorrect or out-of-date data in most cases would lead to an unwanted behavior of the system, e.g. malfunctions.

System Landscape Directory quality: According to the official documentation the System Landscape Directory is the primary source for system landscape information. In contrast, most respondents (over 81 percent) stated that all data types except the SAP technical systems as well as the SAP products are only considered in the SLD system insofar as they are of importance for the collaborative processes with some exceptions (see Fig. 2). The SAP products and technical systems are an exception particularly due to the automatic insertion and update of the components in an available SLD system. Apart from completeness in terms of elements stored within SAP PI another important aspect is the completeness of a specific element, i.e., to which extent corresponding attributes are preset with values in SAP PI. In all cases more than 74 percent of the respondents agreed to 'elements are complete' or 'elements are complete with some exceptions' (see Fig. 3). In the case of SAP products this value even achieves 100 percent. Nevertheless, third party systems, databases as well as third party software products are partly stated as commonly incomplete.

The assessment of the correctness (cf. Fig. 4) reveals that concerning all data types more than 78 percent agreed data to be either accurate or accurate with some exceptions. In the case of SAP products the proportion is 100 percent. This effect may be attributed to the automated insertion and update of data within the SAP product family.

Focus during the analysis of the dimension *actuality* lies on two questions:

- When are changes to the system as well as application landscape taken into account?
- When are decaying data deleted?

Even though the SLD system is considered as the central information source of the system landscape, most respondents of the survey stated that all data types, except SAP technical systems and SAP products are only considered in the SLD system, with some exceptions, when they become important for the collaborative process.

Out-of-date data pose another problem. In particular, this includes elements which are still saved in the SLD system but not used in practice anymore yielding to a faulty reflection of the world. Out of the respondents 76 percent claimed that old data are deleted within one year and shorter and only 24 percent stated a time interval greater than one year. Within a survey conducted by Farwick et al. [6] the respondents reported that an actuality of EA information within weeks (48 per-

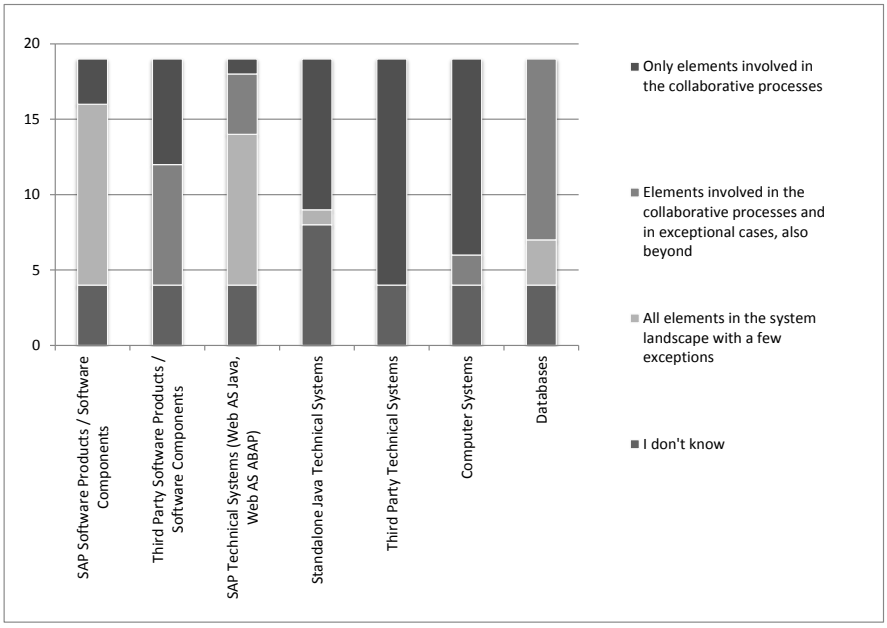


Fig. 2 Elements stored within SLD (n=19)

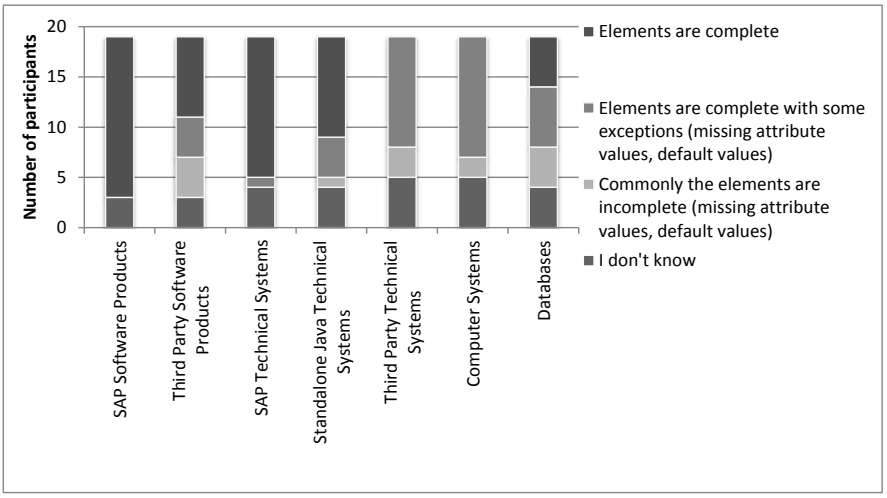


Fig. 3 Completeness of SLD data (n=19)

cent) or up to six months (31 percent) is appropriate. While 76 percent claimed to delete out-of-data data within six months or even earlier, no participant stated an interval shorter than one month.

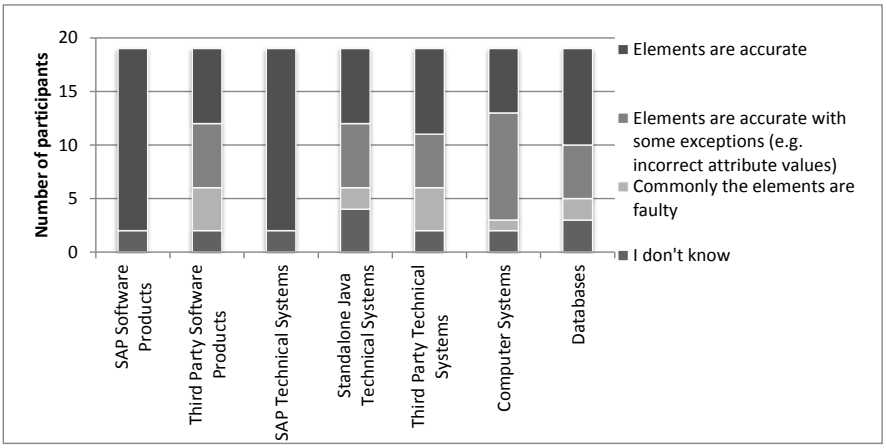


Fig. 4 Correctness of SLD data (n=19)

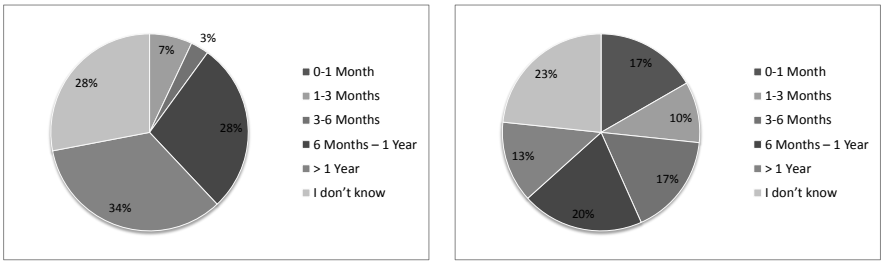
Enterprise Service Builder and Integration Builder quality: Interviews conducted in advance revealed the data of the Enterprise Service Builder as well as the Integration Builder to be nearly error-free and complete due to the mandatory characteristic and consequences to productive system environments in case of errors. Yet, it has to be highlighted that both components only comprise data necessary for the communication processes over the SAP PI system [4]. For instance, unused (legacy) interfaces are not taken into account. This holds true for other data types, e.g. software components, databases and computer systems. As a result, with respect to the quality dimensions (correctness, completeness, actuality), *actuality* was further investigated in our online questionnaire.

Analogous to the SLD system the time to delete is of relevance in order to assess the problem of out-of-date data and the resulting errors (see Fig. 5(a) and Fig. 5(b)). Unfortunately, with respect to the *Enterprise Service Repository* only 15 percent of the respondents reported an interval shorter than six months. In contrast to the average deletion time within the Enterprise Service Builder 56 percent reported a time interval shorter than 6 months for the Integration Builder. 21 percent even agreed to 0-1 month.

4 Effects on the EA Model Quality

Putting our findings in context towards developing and evaluating an automated EA documentation process previous data quality consideration gives a first impression to which extent the generated EA models are affected by the underlying data quality.

Completeness of EA Models: Previous investigations show the information content of EA models is limited to elements used within communication processes. Any



(a) Enterprise Service Builder (b) Integration Builder

Fig. 5 Survey results: Deletion interval (n=19)

information beyond is commonly abstracted even though the SLD officially is meant to be the central information source about the system landscape.

Correctness of EA Models: The previous analysis reveals data of SAP PI systems seem to be correct in most cases especially concerning data stored in the *Enterprise Service Repository* and *Integration Builder*. Accordingly, this also applies to the EA information derived from the data.

Actuality of EA Models: Orphaned data in the SAP PI system pose a problem. In particular, this includes elements which are still saved in the SAP PI system but not used in practice anymore. Consequently, extracted EA information paints a misleading picture not allowing to make appropriate decisions. In contrast to a manual collection in which orphanded data are probably filtered out by the individuals responsible such human filters are dropped within an automation. Losses in quality have to be offset elsewhere.

5 Conclusion and Outlook

This paper gives a first impression which impact data quality aspects of productive IT systems have on automated EA documentation endeavors. Survey results presented show the majority of data in productive SAP PI systems seems to be accurate (complete and correct). Combined with results of our recent study [4], an SAP PI system seems to be 1) suitable and 2) a reasonable starting point for an automated EA documentation endeavor. However, this statement cannot be generalized and further research of productive IT environments is necessary to show the real potential of an automated EA documentation.

Our current efforts are centered around a particular ESB system, namely SAP PI. Further research could integrate Configuration Management Databases (CMDBs) and infrastructure monitoring tools. Both could be very beneficial for impact analyses and strategic planning. Our long term hypothesis is 'when information is gath-

ered from upper layers, e.g. business processes and capabilities, more unstructured information is to expect which would have an impact on 1) the data quality and 2) on the automation potential for EA documentation. With this in mind, a closer look at the field of process mining could be worthwhile.

References

1. A8402 ANSI/ASQC: Quality Management and Quality Assurance - Vocabulary. American Society for Quality Control (1994)
2. A8402 ANSI/ASQC: Quality Management and Quality Assurance - Vocabulary. American Society for Quality Control (1994)
3. Buckl, S., Matthes, F., Roth, S., Schulz, C., Schweda, C.M.: A conceptual framework for enterprise architecture design. In: W. Aalst, J. Mylopoulos, N.M. Sadeh, M.J. Shaw, C. Szyperski, E. Proper, M.M. Lankhorst, M. Schnherr, J. Barjis, S. Overbeek (eds.) Trends in Enterprise Architecture Research, *Lecture Notes in Business Information Processing*, vol. 70, pp. 44–56. Springer Berlin Heidelberg (2010)
4. Buschle, M., Ekstedt, M., Grunow, S., Hauder, M., Roth, S., Matthes, F.: Automating enterprise architecture documentation using models of an enterprise service bus. In: Americas Conference on Information Systems (AMCIS) (2012). (to appear)
5. Buschle, M., Holm, H., Sommestad, T., Ekstedt, M., Shahzad, K.: A Tool for automatic Enterprise Architecture modeling. In: CAISE11 Forum (2011)
6. Farwick, M., Agreiter, B., Breu, R., Ryll, S., Voges, K., Hanschke, I.: Requirements for automated enterprise architecture model maintenance - a requirements analysis based on a literature review and an exploratory survey. In: ICEIS (4), pp. 325–337 (2011)
7. Farwick, M., Agreiter, B., Ryll, S., Voges, K., Hanschke, I., Breu, R.: Automation Processes for Enterprise Architecture Management. In: Trends in Enterprise Architecture Research (TEAR). Helsinki (2011)
8. Glass, R.: Building Quality Software, first edn. Prentice Hall (1991)
9. Hauder, M., Roth, S., Matthes, F.: Challanges for automated enterprise architecture documentation. In: Trends in Enterprise Architecture Research (TEAR). (in submission) (2012)
10. Hauder, M., Roth, S., Schulz, C.: Generating dynamic cross-organizational process visualizations through abstract view model pattern matching. In: Architecture Modeling for the Future Internet enabled Enterprise AMFinE (2012)
11. Kahn, B., Strong, D., Wang, R.: Information quality benchmarks: Product and service performance. Communications of the ACM pp. 184–192 (2002)
12. Kan, S.: Metrics and Models in Software Quality Engineering. Addison-Wesley Longman (2002)
13. Moser, C., Junginger, S., Brückmann, M., Schöne, K.M.: Some Process Patterns for Enterprise Architecture Management. In: Strategy, pp. 19–30 (2009). URL `http://subs.emis.de/LNI/Proceedings/Proceedings150/8.pdf`
14. Nicolescu, V., Funk, B., Niemeyer, P., Heiler, M., Wittges, H., Morandell, T., Visintin, F., Stegemann, B.K., Kienegger, H.: Praxishandbuch SAP NetWeaver PI - Entwicklung, second edn. SAP Press, Bonn (2009)
15. Reeves, C.A., Bednar, D.A.: Defining quality: Alternatives and implications. Academy of Management Review **19**, 419–445 (1994)
16. Ross, J.W., Weill, P., Robertson, D.C.: Enterprise Architecture as Strategy. Harvard Business School Press, Boston, MA, USA (2006)
17. Schaub, M., Matthes, F., Roth, S.: Towards a Conceptual Framework for Interactive Enterprise Architecture Management Visualizations. In: Modellierung (2012)
18. Spewak, S., Hill, S.C.: Enterprise Architecture Planning: Developing a Blueprint for Data, Applications, and Technology, second edn. John Wiley & Sons, London (1993)

19. Steen, M., Strating, P., Lankhorst, M., Doest, H., Eugenia, I.M.E.: Service-Oriented Enterprise Architecture. Idea Group Publishing (205)
20. The Open Group: TOGAF Version 9 – A Manual, 9th edn. Van Haren Publishing (2009)
21. Wang, R.W., Strong, D.M.: Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems **12**, 5–30 (1996)
22. van't Wout, J., Waage, M., Hartman, H., Stahlecker, M., Hofman, A.: The Integrated Architecture Framework Explained: Why, What, How, first edn. Springer, Berlin (2010)
23. Zachman, J.A.: A framework for information systems architecture. IBM Syst. J. **26**(3), 276–292 (1987). URL `http://portal.acm.org/citation.cfm?id=33596`