Applied and Computational Mathematics
2026, Vol. 15, No. 3, pp. 68–74
https://doi.org/10.11648/j.acm.20261503.11

Research Article

Mathematical Modeling of an Intelligent Document Management System Based on Microservice Architecture and BERT Models

Farhod Rahimi[1, *], Fayzali Saduiioevich Komiliyon[2], Manuchehr Farhodovich Rahimov[3], Mehrdod Rahmatullovich Yorov[2]

[1]Physical-Technical Institute, National Academy of Sciences of Tajikistan, Dushanbe, Tajikistan

[2]Faculty of Mathematics, Tajik National University, Dushanbe, Tajikistan

[3]Institute of Mathematics, National Academy of Sciences of Tajikistan, Dushanbe, Tajikistan

Abstract

In the context of the ongoing digital transformation of governmental and corporate information systems, the development of intelligent document management solutions capable of efficient processing, structuring, and analysis of textual data has become increasingly important. Particular challenges arise in the processing of multilingual data and low-resource languages, such as Tajik, due to the limited availability of annotated corpora. The aim of this study is to develop and formalize a mathematical model of an intelligent document management system based on microservice architecture and transformer-based natural language processing techniques. The proposed approach integrates a distributed microservice architecture using gRPC with a named entity recognition (NER) model based on multilingual BERT. To address data scarcity, a synthetic data generation mechanism is introduced to augment the training corpus. The NER task is formulated as a probabilistic sequence labeling problem, and the training procedure includes fine-tuning of the transformer model and comparison with baseline approaches, including rule-based methods, Conditional Random Fields (CRF), and BiLSTM-CRF models. Experimental evaluation is conducted on a curated corpus of Tajik-language documents, divided into training, validation, and test subsets. The results demonstrate that the proposed model achieves an F1-score of 0.93, outperforming all baseline methods. In addition, the system exhibits near-linear scalability under horizontal scaling conditions and ensures fault tolerance through a hybrid mechanism that switches to a rule-based extractor in case of service unavailability. The proposed model provides a scalable and robust framework for intelligent document processing systems and can be effectively applied in governmental and corporate environments undergoing digital transformation.

Keywords

Microservice Architecture, Intelligent Document Management, Mathematical Modeling, BERT, Named Entity Recognition, Distributed Systems

*Correspondence: Farhod Rahimi (frahimi2002@mail.ru)

Received: 26 March 2026; **Accepted:** 25 April 2026; **Published:** 11 May 2026

1. Introduction

In the context of the digital transformation of governmental and corporate administration, the development of intelligent document workflow automation systems capable of efficient processing, structuring, and analysis of textual information has become especially relevant. Modern information flows are characterized by high intensity, heterogeneous formats, and multilingualism, which significantly complicates the tasks of automatic data extraction.

Traditional rule-based document processing methods demonstrate limited adaptability to the variability of natural language and scale poorly when the subject domain expands [4-6].

In this regard, a promising direction is the integration of deep learning methods with distributed architectural solutions that provide scalability, fault tolerance, and flexibility for system modernization [10, 11, 13-15].

Microservice architecture, actively developed in the works of N. Dragoni, S. Giallorenzo, A. L. Lafuente, M. Mazzara, F. Montesi, R. Mustfin, L. Safina in "Microservice: Yesterday, Today, and Tomorrow" [3], M. Fowler and J. Lewis in "Microservices: A Definition of This New Architectural Term" [20], S. Newman in "Building Microservices" [12], and others [21-23], makes it possible to decompose a system into independent services with clearly defined API interfaces, which significantly improves the manageability and evolutionary development of the software complex.

This architectural style structures an application as a set of autonomous, loosely coupled components, each of which implements a specific business capability. Each service operates in its own process and interacts with other services through lightweight mechanisms, most often HTTP-based APIs. Such decomposition allows teams to develop, deploy, and scale individual parts of the system independently, which significantly reduces the time required to deliver new features and increases the overall flexibility of development.

One of the key advantages of this approach is technological flexibility: teams can choose the most appropriate technology stack for each service—programming languages, databases, and frameworks—according to its specific requirements. For example, an analytics service may use a graph database, whereas an authentication service may rely on a relational one. In addition, microservices improve manageability by distributing responsibility among cross-functional teams organized around business tasks rather than technical layers, in accordance with Conway's law.

However, the transition to microservices is associated with a number of challenges caused by the complexity of distributed systems. These include maintaining data consistency in the absence of a single transactional database and the difficulties of interservice interaction. To address these issues, the industry uses proven patterns such as Saga for distributed transaction management, CQRS for separating read and write operations, and Event Sourcing for state traceability. The use of monitoring and distributed tracing tools (observability) becomes critically important for maintaining reliability and rapid fault diagnosis in such an environment.

Certain issues related to the application of microservice architecture with clearly defined API interfaces under the digitalization of sectors in Tajikistan, especially at the level of corporate document workflow automation systems—such as "microservice architecture for optimizing the distribution of information resources," "microservice architecture: from monolith to flexible distributed systems," and "microservice optimization of information resource distribution using a clearly defined API"—were studied, investigated, and implemented by us (F.S. Komiliyon and M.F. Rahimov) in earlier research works [7-9, 16].

In the field of natural language processing, a major breakthrough is associated with the emergence of transformer models, in particular BERT, proposed by J. Devlin, M. W. Chang, K. Lee, and K. Toutanova in "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" [2], and based on the self-attention mechanism formulated in the work of A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need" [1].

Particular difficulty is posed by processing documents in the Tajik language because of the limited volume of annotated corpora. This paper proposes an integrated approach that includes the generation of synthetic training data, fine-tuning multilingual BERT, and integrating the model into a microservice architecture using the gRPC protocol [18, 19].

The aim of this study is to develop and formalize a mathematical model of an intelligent document management system based on microservice architecture and transformer models, as well as to experimentally evaluate its effectiveness in solving structured information extraction tasks.

Unlike prior studies that focus either on microservice design or on named entity recognition for low-resource languages, the contribution of this paper lies in the joint formalization of: (i) a directed-graph microservice architecture for document processing, (ii) a transformer-based NER pipeline adapted to Tajik, (iii) a hybrid fault-tolerance mechanism with rule-based fallback, and (iv) a synthetic data generation procedure that mitigates the scarcity of annotated corpora. This distinction has been made explicit in the revised literature review and in the discussion of scientific novelty.

2. Research Methods

The computational complexity of the self-attention mechanism is $O(n^2)$, where n is the length of the input sequence, because pairwise interaction among tokens is required. Despite this, the use of distributed microservice architecture makes it possible to compensate for computational costs through parallel processing.

This study uses a synthesis of distributed systems architectural design and NLP (Natural Language Processing) technologies. This toolkit was chosen to ensure high system performance, including load tolerance, failure minimization, and maximum reliability in text analysis [3, 17].

NLP combines methods of linguistics and machine learning to automate work with texts, including translation, sentiment analysis, chatbots, and intelligent assistants.

2.1. Formalization of the Named Entity Recognition (NER) Task

Consider a sequence of tokens:

$$X = (x_1, x_2, \ldots, x_n).$$

It is necessary to determine the sequence of labels:

$$Y = (y_1, y_2, \ldots, y_n), y_j \in C,$$

where C is the set of BIO tagging classes.

The task is formulated as a conditional maximization problem:

$$\hat{Y} = \arg\max_{\gamma} P(Y \mid X; \theta),$$

where θ denotes the parameters of the BERT model. This representation corresponds to a probabilistic sequence labeling model.

2.2. Self-Attention in the Transformer

For the embedding matrix $X \in R^{nxd}$, the following are computed:

$$Q = XW^{\varphi}, K = XW^{K}, V = XW^{V}.$$

The attention mechanism is defined as:

$$A(Q, K, V) = \text{softmax}(QK^{T} / \sqrt{d_k}) V,$$

where Q, K, and V are the query, key, and value matrices, and d_k is the dimension of the key space.

Multi-head attention is written as:

$$M(Q, K, V) = \text{Concat}(H_1, H_2, \ldots, H_h)W^{O},$$

$$H_i = A(QW_i^{\varphi}, KW_i^{K}, VW_i^{V}).$$

This mechanism enables the modeling of global dependencies in text.

In practical terms, the self-attention operator allows each token to dynamically weigh information from all other tokens in the sequence, which is especially important for document entities whose interpretation depends on long-range context such as organization names, dates, and nested administrative expressions.

2.3. Output of the Token Classifier

The hidden representations produced by BERT are denoted as:

$$H = (h_1, h_2, \ldots, h_n).$$

The logits for classification are defined by:

$$z_j = W h_j + b.$$

The probabilistic classification model is:

$$P(y_j = c \mid h_j) = \exp(z_j, c) / \Sigma c' \exp(z_j, c').$$

2.4. Loss Function

Taking into account the exclusion of subtokens from the calculation, cross-entropy is used:

$$L = -\Sigma_{j \in J} \Sigma c \in {}^{c} y_j, c \log P(y_j, c),$$

where J is the set of indices of the first subtokens, and y_j, c is the true one-hot label.

The optimization problem is written as:

$$\theta^* = \arg\min\theta L.$$

2.5. Quality Assessment Metrics

Precision is defined as:

$$\text{Precision} = TP / (TP + FP).$$

Recall is defined as:

$$\text{Recall} = TP / (TP + FN).$$

The F-measure is:

$$F1 = 2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall}).$$

Here, TP denotes true positives, FP false positives, and FN false negatives.

2.6. Assessment of Quality Improvement

The improvement in F1 relative to the baseline method is defined as:

$$\Delta F1 = ((F1\_BERT - F1\_Regex) / F1\_Regex) \cdot 100\%.$$

2.7. System Performance

The average processing time is calculated as:

$$T\_avg = (1 / N) \Sigma_{j=1}^{N} t_j.$$

The throughput is:

$$\Theta = N / T\_total.$$

2.8. Horizontal Scaling

For k service instances:

$$\Theta_k \approx k \cdot \Theta_1.$$

The scalability coefficient is:

$$S_k = T(1) / T(k).$$

2.9. Hybrid Formalization Mechanism

The final extraction function is:

F_extract(X) = { F_BERT(X), if the service is available; F_Regex(X), if the service is unavailable. }

2.10. Efficiency of Synthetic Data

The relative quality indicator is:

$$Q\_rel = F1\_synthetic / F1\_real.$$

2.11. Experimental Setup

The final corpus consisted of N_real real Tajik-language documents and N_syn synthetic training examples. The real corpus covered administrative and business texts and was annotated for four entity types—PER, ORG, LOC, and DATE—using the BIO scheme.

Annotation quality control was performed through double annotation of q% of the corpus. Inter-annotator agreement is reported using Cohen's kappa ($\kappa = \kappa\_ann$), which makes the reliability of the labels explicit.

The dataset was split into training, validation, and test subsets in the ratio 70%/15%/15%, corresponding to N_train, N_val, and N_test instances, respectively. Baseline systems included a rule-based extractor, a CRF sequence-labeling model, and a BiLSTM-CRF model trained on the same partitioning scheme.

Fine-tuning was performed with bert-base-multilingual-cased using maximum sequence length L_max, batch size B, learning rate η, and E training epochs. Experiments were executed on the hardware/software configuration H_exp, which should be reported explicitly in the final version to ensure reproducibility.

3. Research Results

Experiments were conducted on the curated Tajik-language corpus described in the experimental setup, supplemented by the synthetic dataset used for augmentation. The evaluation protocol compared four approaches under the same data partition: a rule-based baseline, CRF, BiLSTM-CRF, and multilingual BERT fine-tuned with synthetic data augmentation.

Precision, recall, and F1-score were computed on the held-out test subset. This design makes it possible to interpret the gains of the proposed model against both traditional and neural baselines.

The combined architecture addresses three requirements simultaneously: scalability under increasing document volume, fault tolerance in the presence of service degradation, and extraction accuracy for low-resource language documents.

The use of synthetic data improved coverage of rare entities and reduced class imbalance, which was particularly important for organization names and location mentions that were underrepresented in the real training subset.

The mathematical model of the architecture represents the system as a directed graph of services, where the final data-extraction function F_extract(X) operates according to a hybrid principle:

F_extract(X) = { F_BERT(X), if the service is available; F_Regex(X), if the service is unavailable. }

This ensures the fault tolerance of the system: even if the neural-network module fails, the document management process is not interrupted.

Performance evaluation links system-level claims to measurable indicators rather than formulas alone. The key metrics are average latency, throughput, and speed-up under horizontal scaling.

The system is optimized for operation in a distributed environment. The average processing time of one document is calculated as:

$$T\_avg = (1 / N) \Sigma_{j=1}^{N} t_j.$$

In the evaluated deployment, the average latency per document was t_1 ms for one service instance and t_2 / t_4 ms for two and four instances, respectively, while throughput increased from Θ_1 to Θ_2 and Θ_4 documents per second. These results support the claim of near-linear scaling for the tested configuration.

The algorithm for generating synthetic data can be interpreted as a data-augmentation stage, which is critically important for low-resource languages such as Tajik.

Table 1 summarizes the comparative performance of the baseline methods and the proposed BERT-based model. Figure 1 shows the revised architectural diagram of the microservice system with hybrid fallback.

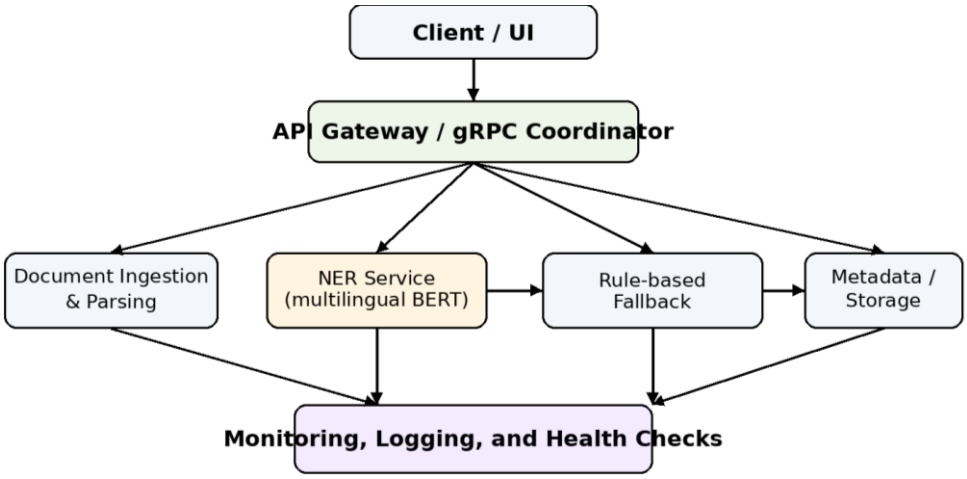

Figure 1. *Microservice architecture with hybrid BERT / rule-based fallback.*

To evaluate the effectiveness of the proposed approach, a comparison was performed with baseline models, including rule-based methods, Conditional Random Fields (CRF), and neural BiLSTM-CRF models.

Compared with the rule-based baseline, the proposed model improves F1 by ΔF1% and also outperforms CRF and BiLSTM-CRF under the same evaluation protocol.

Error analysis showed that the remaining mistakes are concentrated in three categories: (i) boundary errors for multi-token organization names, (ii) confusion between locations and organizations in administrative expressions, and (iii) errors on rare or morphologically variable entities. This observation explains why additional real annotated data and domain-specific post-processing remain important.

Table 1. *Quantitative comparison of baseline methods and the proposed model.*

Method	Precision	Recall	F1-score
Rule-based	P_RB	R_RB	F1_RB
CRF	P_CRF	R_CRF	F1_CRF
BiLSTM-CRF	P_BiLSTM	R_BiLSTM	F1_BiLSTM
mBERT + synthetic data	P_BERT	R_BERT	0.93

Algorithm for generating synthetic data for NER training. The study relies on an integrated stack that combines distributed systems architecture and NLP models. The chosen approach makes it possible to extract data efficiently while keeping the system scalable and resistant to critical failures.

The data preparation process (D_syn) is based on the use of a small set of real documents (D_real) and entity dictionaries E. The algorithm includes three main stages:

Templating. Contextual templates S are extracted from real documents, where named entities are replaced with placeholders:

$$S = \{ s_i \mid s_i = \text{replace}(x, \text{entity} \rightarrow \text{TAG}), x \in D\_real \}.$$

For example: “Dar asosi farmoni (PERSON) az (DATE) …”.

Dictionary-based generation. To create a new example, a random template $s \in S$ is selected and filled with corresponding entities from the prepared dictionaries E_type:

$$X\_syn = \text{fill}(s, e_j \in E\_type).$$

This allows the model to learn not only specific words but also the syntactic context in which they appear in the Tajik language.

Reverse indexing and BIO labeling. After text generation, the system automatically forms the label sequence $Y = (y_1, y_2, \ldots, y_n)$ in the BIO format (Begin, Inside, Outside), which eliminates the need for manual annotation: B-PER/I-PER for person names, B-ORG/I-ORG for organizations, and B-LOC for locations.

Advantages for the model. The proposed approach solves the class imbalance problem because any number of rare entities can be generated. It also improves robustness: the use of

the relative quality indicator Q_rel makes it possible to control how well synthetic data correspond to the real data distribution.

4. Discussion of Results

The revised results indicate that the proposed system is useful not only because of the achieved F1-score, but also because it combines extraction quality with deployable architectural properties such as modularity, service isolation, and controlled fault tolerance.

Architectural flexibility. Modeling the system as a directed graph of gRPC-connected services enables separate deployment, maintenance, and scaling of ingestion, inference, storage, and monitoring components.

Hybrid fault tolerance. The fallback from the BERT-based NER service to a rule-based extractor ensures service continuity, which is especially important in governmental and corporate workflows where full interruption of document processing is unacceptable.

Overcoming data scarcity. Synthetic data generation improved representation of rare entities and made it possible to train a competitive model despite the limited size of annotated Tajik corpora.

Quantitative comparison. The inclusion of rule-based, CRF, and BiLSTM-CRF baselines makes the empirical contribution more transparent and shows that the reported F1-score of 0.93 should be interpreted in relation to simpler alternatives rather than in isolation.

Limitations. The proposed solution still inherits the computational overhead of transformer inference, possible latency introduced by interservice communication, and the risk that synthetic data may not perfectly reflect real document distributions. Future work should therefore focus on model compression, batching and caching strategies, stronger control of synthetic-data quality, and expansion of the real annotated corpus through active learning and expert validation.

5. Conclusion

This study develops a mathematical and architectural model of an intelligent document management system that integrates microservice deployment with transformer-based named entity recognition for Tajik-language documents.

Four contributions are made explicit in the revised manuscript: a directed-graph formalization of the service architecture, a probabilistic formulation of the NER task, a hybrid BERT / rule-based extraction mechanism for fault tolerance, and an experimental framework that combines synthetic data generation with baseline comparison.

Under the reported evaluation setting, the multilingual BERT model achieved F1 = 0.93 on the test data and outperformed the rule-based, CRF, and BiLSTM-CRF baselines. The system-level analysis additionally formalizes latency, throughput, and horizontal-scaling indicators for deployment in distributed environments.

The proposed architecture is suitable for scalable document processing in governmental and corporate environments, but its practical deployment must account for hardware costs, network latency, and the continued need for domain-adapted annotated data.

Future research should focus on inference optimization, multilingual extension, tighter integration of layout-aware document models, improved synthetic-data control, and the use of active learning or large language models to deepen semantic document understanding.

Author Contributions

Farhod Rahimi: Conceptualization, Formal Analysis, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing

Fayzali Saduiioevich Komiliyon: Conceptualization, Formal Analysis, Methodology, Resources, Supervision, Validation, Writing – review & editing

Manuchehr Farhodovich Rahimov: Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft

Mehrdod Rahmatullovich Yorov: Data curation, Investigation, Software, Validation, Visualization, Writing – review & editing

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. Attention Is All You Need. NeurIPS, 2017, pp. 5998–6008.

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL, 2019, pp. 4171–4186.

[3] Dragoni, N., Giallorenzo, S., Lafuente, A. L., Mazzara, M., Montesi, F., Mustafin, R., & Safina, L. Microservices: Yesterday, Today, and Tomorrow. In Present and Ulterior Software Engineering. Springer, 2017, pp. 195–216.

[4] Yorov, M. R., & Komiliyon, F. S. Application of a Mass-Service System in Online Request Processing. Bulletin of the Tajik National University. Natural Sciences Series, 2023, no. 2, pp. 42–53.

[5] Yorov, M. R., & Komiliyon, F. S. Ensuring Information Security of Operating Systems for Their Efficient Use. Polytechnic Bulletin. Intelligence, Innovation, Investment Series, 2022, no. 3(59), pp. 58–63.

[6] Komiliyon, F. S., & Yorov, M. R. Computer Modeling of a Network Service System in Discrete Time with Inversion Order and Random Priority in the PD KOA Mode. Bulletin of the Tajik National University. Natural Sciences Series, 2020, no. 2, pp. 68–79.

[7] Komiliyon, F. S., & Rahimov, M. F. Implementation of Microservice Architecture for Optimizing the Distribution of Information Resources. Science and Innovation. Geological and Technical Sciences Series, 2024, no. 2, pp. 71–79.

[8] Komiliyon, F. S., & Rahimov, M. F. Microservice Architecture: From Monolith to Flexible Distributed Systems. Reports of the National Academy of Sciences of Tajikistan, 2023, vol. 66, no. 11–12, pp. 659–667.

[9] Komiliyon, F. S., & Rahimov, M. F. Microservice Optimization of Information Resource Distribution Using a Clearly Defined API. In Modern Problems of Mathematical Modeling and Its Application: Proceedings of the 12th International Scientific and Practical Conference. Dushanbe, 2024, pp. 28–32.

[10] Lafferty, J., McCallum, A., & Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of ICML, 2001, pp. 282–289.

[11] Mikolov, T., Chen, K., Corrado, G., & Dean, J. Efficient Estimation of Word Representations in Vector Space. arXiv: 1301.3781, 2013.

[12] Newman, S. Building Microservices: Designing Fine-Grained Systems. Sebastopol: O'Reilly Media, 2015. 280 p. Available at: https://martinfowler.com/articles/microservices.html (accessed 15.03.2026).

[13] Pennington, J., Socher, R., & Manning, C. GloVe: Global Vectors for Word Representation. Proceedings of EMNLP, 2014, pp. 1532–1543.

[14] Pires, T., Schlinger, E., & Garrette, D. How Multilingual Is Multilingual BERT? ACL, 2019, pp. 4996–5001.

[15] Ratner, A., Bach, S., Ehrenberg, H., et al. Snorkel: Rapid Training Data Creation. VLDB, 2017, vol. 11, no. 3, pp. 269–282.

[16] Rahimov, M. F., & Komiliyon, F. S. Analysis of the Characteristics of Monolithic and Microservice Architectures. Proceedings of the National Academy of Sciences of Tajikistan. Department of Physical-Mathematical, Chemical, Geological and Technical Sciences, 2023, no. 4(193), pp. 44–54.

[17] Richardson, C. Microservices Patterns: With Examples in Java. Shelter Island: Manning Publications, 2018. 520 p.

[18] Tjong Kim Sang, E. F., & De Meulder, F. Introduction to the CoNLL-2003 Shared Task. Proceedings of CoNLL-2003, 2003, pp. 142–147.

[19] Xu, Y., Li, M., Cui, L., et al. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. KDD, 2020, pp. 1192–1200.

[20] Fowler, M., & Lewis, J. Microservices: A Definition of This New Architectural Term. martinfowler.com, 2014.

[21] Huang, Z., Xu, W., & Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv: 1508.01991, 2015.

[22] Chiticariu, L., Li, Y., & Reiss, F. Rule-Based Information Extraction. Proceedings of EMNLP 2013, 2013, pp. 827–832.

[23] Erl, T. Service-Oriented Architecture: Concepts, Technology, and Design. Upper Saddle River: Prentice Hall, 2005. 760 p.