

Organic Data:

Ein sicheres, dezentralisiertes Big Data Konzept

Sebastian von Mammen, Carsten Grenz, Jörg Hähner Sabine Timpf
Organic Computing Geoinformatik
Universität Augsburg Universität Augsburg
Eichleitnerstr. 30 Alter Postweg 118
86159 Augsburg 86159 Augsburg
{sebastian.von.mammen,carsten.grenz,joerg.haehner}
@informatik.uni-augsburg.de
sabine.timpf@geo.uni-augsburg.de

Daniel Loebenberger Stefan Mandl, Oleksandr Kozachuk
Dept. for Computer Security EXASOL AG
Bonn-Aachen Int. Center for IT Neumeyerstr. 48
Dahlmannstr. 2 90411 Nürnberg
53113 Bonn
daniel@bit.uni-bonn.de
{stefan.mandl,oleksandr.kozachuk}@exasol.com

Abstract: In diesem Papier stellen wir ein Konzept vor für die sichere, dezentralisierte Sammlung und Verwertung großer Datenmengen. Die Kernidee ist eine selbstorganisierte hierarchische Organisation cyber-physikalischer “Organic Data”-Knoten (ODNs), die fähig sind, Daten lokal zu speichern, zu verarbeiten und gezielt unter Nachbarknoten zu kommunizieren. Das resultierende Netzwerk ist eine sich anpassende virtuelle Struktur, in der Daten aggregiert, verarbeitet und gespeichert werden. Der präsentierte Ansatz schließt Business-to-Consumer-Modelle gestützt auf potentiell personenbezogenen Datensätzen explizit mit ein. Neben der allgemeinen Darstellungen des Konzepts und weiterführender Erläuterungen hinsichtlich zugrundeliegender Technologien, werden konkrete Benchmarks für seine Evaluation präsentiert.

1 Einführung

Die massenhafte Verarbeitung und Speicherung von Daten erfordert ein besonderes Augenmerk auf den Schutz der Privatsphäre [BF10, KTGH13]. Bestehende Big Data Anwender wie z.B. Google und Facebook werden diesem Anspruch keinesfalls gerecht. Ihre Geschäftsmodelle und ihre DV-Infrastruktur zielen darauf ab, sämtliche Nutzerdaten zentral und weitestgehend ungefiltert zu horten¹, was beispielsweise die kürzlich offengelegten unautorisierten Zugriffe seitens der amerikanischen und britischen Sicherheits-

¹www.datacenterknowledge.com

behörden in hohem Maß begünstigte [Hol13].

Wir begegnen dieser Herausforderung mit *Organic Data*, einem dezentralen, selbstorganisierenden Ansatz zu Big Data, der dem Missbrauch persönlicher Daten vorbeugt und gleichermaßen effizient und skalierbar ist. Anstatt die Daten verteilter Quellen zentral zu aggregieren, werden sie von lokalen Netzwerkknoten gezielt angefragt und verarbeitet, um dann gegebenenfalls als verallgemeinertes Faktum einer Wissensbasis hinzu gefügt oder als erkanntes Ereignis an andere Systemknoten weitergemeldet zu werden. Die Empfänger verfahren genauso, nur auf einem höheren Abstraktionsniveau, sodass Wissen über das beobachtete System umfassend und in hohem Maße abstrahiert gesammelt wird. Der grundlegenden Frage der Partitionierung der Datenbestände in einem verteilten Datenbanksystem wird durch die cyber-physikalische Netzwerkstruktur begegnet; sie ergibt sich, wie im nächsten Abschnitt genauer erläutert, automatisch durch das Zusammenspiel von (authorisierten) Benutzeranfragen, der initialen Gewinnung von Daten und ihrer weiterführenden Verarbeitung und Kommunikation.

Organic Computing Algorithmen [MSSU11] sollen die Topologie und die Kommunikation im Netzwerk "selbst"-organisieren, um sie einerseits den dynamischen Abstraktionen anzupassen und andererseits, um zu jedem Zeitpunkt Redundanz und Ressourcenverbrauch der Netzwerkknoten zu optimieren. Die Knoten des Organic Data Netzwerk sammeln, lernen, kanalisieren Daten und bedienen außerdem lokale, hochperformante Datenbanken.

Aufgrund der dezentralen Organisation von Organic Data muss im Anwendungskontext speziell auf Datensicherheit, Visualisierung und Benutzerfreundlichkeit eingegangen werden. Nutzung von Verschlüsselungstechniken, Authentisierungs- und Anonymisierungsverfahren, sowie die Verteilung auf mehrere Knoten und die möglicherweise vielschichtige Verarbeitung sollen unberechtigte Zugriffe auf die Daten verhindern. Die eingesetzte Sicherheitstechnik soll insbesondere Datenschutz auf hohem und quantifizierbarem Niveau gewährleisten: *privacy by design*. Ein Visual Analytics System soll es ermöglichen, Informationen der Netzwerkknoten einzusehen und höher- und tieferliegende Ebenen der Netzwerk- und Abstraktionshierarchie interaktiv zu steuern. Eine Herausforderung dabei liegt in der notwendigen dynamischen Anpassung des der Visualisierung zugrundeliegenden Datenmodells an den Abstraktionsgrad. Die Interaktion soll auf intelligente Art und Weise eine explorative Analyse der Daten unterstützen, ohne den Datenschutz zu verletzen.

Weiterhin ist dieser Beitrag wie folgt gegliedert. In Sektion 3 nehmen wir Bezug auf den Stand der Forschung aus verschiedenen, für Organic Data relevanten Blickwinkeln und erläutern das Wechselspiel der benannten Aspekte. In Sektion 4 werden Möglichkeiten dargelegt, um konkrete Implementierungen des Organic Data Ansatzes zu evaluieren. Wir schließen den Beitrag mit einer Zusammenfassung in Sektion 5.

2 Datenfluss im Organic Data Netzwerk

Anstatt wie bei herkömmlichen Big Data Ansätzen, Quelldaten zentral zu aggregieren, sollen sie durch ein Netzwerk von *Organic Data Nodes* (ODNs) geschleust werden. Organic

Data Nodes bieten eine Laufzeitumgebung für Softwareagenten, die die Kommunikation im Netzwerk koordinieren, empfangene Daten filtern, abstrahieren und lokal speichern. Dabei soll die Vertraulichkeit und Anonymität der Daten an jedem Punkt des Systems gewährleistet werden. Zugriffe auf die lokalen Daten verschiedener ODNs sollen aufgrund eingehender und ausgehender Datenströme erteilt werden. Wir wollen untersuchen, inwieweit diese Klassifizierung automatisch oder durch möglicherweise verteilte Autoritäten erfolgen muss.

2.1 Selbstorganisation nach Lernzielen

Die ODNs sollen sich mittels entsprechender Methoden des Organic Computing selbstorganisieren, sodass sich der Datenfluss aus der jeweiligen Problemstellung ergibt. Unser algorithmischer Entwurf sieht vor, dass eine bestimmte Anzahl von ODNs die Informationen von einer großen Anzahl an Quellen bezieht, verarbeitet und an einige wenige Senken weitergibt. Der relative Informationsgewinn entscheidet, ob der Datenfluss zwischen Knotenpaaren zu- oder abnimmt. Durch diesen Mechanismus kann der Anwender die Problemstellung definieren, indem er beschreibt, welche Zustände oder Prozesse ihn interessieren, also großen subjektiven Informationsgewinn versprechen (bspw. eine schnell wachsende Anzahl von Verkehrsteilnehmern pro Flächeneinheit). Der Anwender kann diese Bewertungen kontinuierlich verändern, um Abstraktion und Datenfluss im Organic Data Netzwerk zu verfeinern. Auch kann sich das Netzwerk durch dieses Prinzip ständig an neuartige Beobachtungen (Anomalien) oder Veränderungen des Netzwerks (bspw. durch Ausfall oder Hinzunahme von Quellen, Senken oder ODNs) anpassen.

Die resultierende Kommunikationstopologie des Organic Data Netzwerks korreliert mit den Lern- bzw. Abstraktionsprozessen. Entsprechend spiegelt sie sich auch in der lokal umgesetzten Datenpersistenz wider: Jeder ODN speichert, je nach Privacy-Modell und Anwendungsdomäne, einkommende Daten (auf erster Knotenebene u.a. Rohdaten) und/oder abstrahierte, ausgehende Daten. Ein in hohem Grad verteilter, hierarchisch organisierter Datenbestand und der Einsatz moderner kryptographischer Verfahren erschwert nicht nur den unerlaubten Zugriff Dritter, sondern ermöglicht auch die nahtlose Skalierung hochperformanter In-Memory Datenbanksysteme.

2.2 Hierarchischer Zugriff

Der effiziente und authentifizierte Zugriff auf den gesamten Datenbestand des Organic Data Netzwerks - von hochgradig abstrahierten Informationen bis hinab zu Rohdaten (falls vorhanden) - erfolgt durch das Zusammenspiel dreier Komponenten.

Datenbankorganisation Anfragen, die auf höherer Ebene nicht bedient werden können, werden über die unmittelbaren ODN-Quellen abgewickelt.

Authentisierungsmechanismen Zugriffe werden bezüglich der Datengranularität (bis zu

Rohdateneinsichtnahme) und der Vollständigkeit (bis zur Beschreibung des Gesamtsystems) separat authentifiziert. Dazu soll eine globale Schlüsselhierarchie eingeführt werden, die den Zugriff auf die einzelnen Datenblöcke koordiniert.

Visual Analytics Methoden Eine hierarchische Visualisierung der ODNs erlaubt es, (vertikal) in einzelne Knoten einzutauchen und dadurch die Granularität der abgebildeten Daten zu erhöhen. Auf gleichbleibender Abstraktionsebene ermöglicht es eine graphbasierte Darstellung, die Zusammenhänge der abstrahierten Daten sowie der ODNs (horizontal) abzubilden.

3 Stand der Forschung

Organic Data vereint Aspekte aus mehreren Forschungsbereichen, um ein integriertes, skalierbares System zur Verfügung zu stellen. Organic Computing Ansätze dienen als algorithmische Grundlage der selbständigen Organisation von Netzwerktopologie und Datenbeständen. Ansätze aus dem Bereich verteiltes, hierarchisches Lernen werden gebraucht, um Daten in einem dezentralen System verarbeiten und verwalten zu können. Mit der dezentralen, hierarchischen Organisation gehen darüberhinaus besondere Anforderungen an den Schutz der Daten, an die Skalierbarkeit der Speicherung sowie die Visualisierung und Anwenderfreundlichkeit einher.

3.1 Selbstorganisierende Netzwerklösungen

Die Selbstorganisation von Computernetzwerken steht im Mittelpunkt verschiedener Organic Computing Lösungen [MSSU11], welche häufig Natur-inspirierten Algorithmen verwenden. Diese Lösungen sind skalierbar, robust und bieten eine hohe Verfügbarkeit des Systems, indem sie ein globales Management zu Gunsten verteilter in-network Managementalgorithmen aufgeben [Dre06].

Im Bereich von vernetzten ad-hoc Netzwerken spielt häufig die Topologiebildung zur Laufzeit eine wichtige Rolle. So wird z.B. das Finden kurzer Wege beim Routing mit Hilfe von virtuellen Ameisen gelöst [GSB02]. Auch die Optimierung von Ablageorten von Daten zur Laufzeit im Netzwerk kann mit Hilfe von Schwarmalgorithmen gelöst werden [MWT11]; so auch in peer-to-peer Systemen [FLMM10]. Des Weiteren können die eingesetzten Netzwerkprotokolle selbst zur Laufzeit dynamisch angepasst werden [TZHMS10, THH10].

Ein Anwendungsbeispiel für selbstorganisierende Netzwerke stellen aktive Smart Camera Systeme dar, die aus Kameras und Sensoren mit unterschiedlicher Ausprägung und Mobilitätseigenschaften bestehen können [GJH⁺12]. Planungsalgorithmen nutzen die Mobilität von Knoten zur Lösung von Aufgaben aus und haben somit direkten Einfluss auf die Netztopologie [WGH11].

3.2 Verteiltes und hierarchisches Lernen

Die ODNs werden Muster im Stil eines Feed-Forward Netzwerks einkommende Daten lernen und die resultierenden Lernhypothesen weiterleiten. Es konnte gezeigt werden, dass ein derart verteilter Lernansatz redundante, unsichere Quelldaten zusammenführen kann, um die Lernhypothese zu verbessern [CS93]. Die parallele Verarbeitung erlaubt außerdem bestimmte (globale) Lösungen schneller zu finden als mit zentralisierten Systemen [SWA08]. Falls das Problem nicht von beliebigen Blickwinkeln aus erlernt werden kann, bzw. falls keine invariante Partitionierung des Suchraums möglich ist, kann man zumindest sicherstellen, dass man nur jene Muster lernt, die über den gesamten Suchraum Gültigkeit besitzen [PH94]. Immens steigende Datenvolumen motivierten die Entwicklung verteilter Lernmethoden. Mittlerweile werden sie wegen ihrer Robustheit und Effizienz v.a. im Kontext drahtloser Sensornetze diskutiert [PKP06].

Die Zusammenführung verteilten Wissens reicht von der Aggregation (gefilterter) Teilergebnisse, über ihre Akkumulation, bis hin zum Lernen auf den bisherigen Lernhypothesen (Meta-Learning oder hierarchisches Lernen) [vMSDJ11, vMSSJ12, vMS13]. Bei numerischen Werten bietet es sich an, Cluster zu lernen, bspw. mittels künstlicher neuronaler Netze [SvMJ10] oder genetischer Algorithmen [SvMJ11]. Regelmäßige Repräsentationen können durch die Verallgemeinerung ihrer Konditionen und die Aggregation ihrer Aktionen zusammengeführt werden [SDv⁺ed].

Jeder ODN wird mittels lokal akkumulierten oder generierten Wissens (Lernhypothesen) den Informationsgehalt einkommender Datenströme bewerten. Diese Bewertungen liegen der Selbstorganisation im Organic Data Netzwerk zugrunde. Außerdem können durch das Traversieren des Organic Data Netzwerks einerseits abstraktere, andererseits umfassendere Lernhypothesen oder Muster generiert werden, die dem Anwender schließlich präsentiert werden.

3.3 Security/Privacy

Die Kommunikation der ODNs soll mit modernen Verfahren Ende-zu-Ende verschlüsselt werden. Das dazu nötige Schlüsselmaterial wird mittels einer Public-Key-Infrastructure (PKI) bereitgestellt. Hierfür eignet sich eine moderne Blockchiffre wie der Advanced Encryption Standard (AES) [fip01]. Dieser wird in gängigen Sicherheitsprotokollen des Netzwerkverkehrs standardmäßig eingesetzt, bspw. IPsec [KS05] oder TLS [DR08, Eas11].

Einem Angreifer muss es unmöglich sein, einen eigenen Knoten in das Netzwerk einzubringen. Auf Netzwerkebene eignen sich hierfür Message-Authentication-Codes wie CBC-MACs oder CMACs [BR05, nis01]. Auch asymmetrische Authentisierungsverfahren, wie der Digital Signature Standard (DSS) [NIS09], können je nach Anforderung gezielt eingesetzt werden. Letztere erfordern eine dedizierte Schlüssel-Infrastruktur, die beispielsweise auf X.509 Zertifikaten beruhen kann, siehe [CSF⁺08].

Die einzelnen Datenquellen dürfen nicht rück-verfolgbar sein. Dies kann beispielsweise dadurch erfolgen, dass anstelle der Identität einer Datenquelle ein randomisierter Hashwert

übermittelt wird. Hierfür können Hash-Verfahren, wie der Secure Hashing Algorithm (SHA-3) [BDPA13], eingesetzt werden. Diese haben die Eigenschaft, dass aus der Berechnung des Hashwerts keine Rückschlüsse über die Eingabe gezogen werden können.

Die Integration der dargelegten Verfahren ist insbesondere wegen der dezentralen und inhärent dynamischen Organisationsstruktur des Organic Data Ansatzes von großem Interesse.

3.4 Skalierbarkeit von Big Data Lösungen

Obwohl aktuelle High-End Ethernet Karten bereits einen Durchsatz von 56 Gigabit/Sekunde bieten (in beide Richtungen) und sich somit der Durchsatz im Netzwerk in großen Schritten dem Durchsatz des Hauptspeichers annähert, skaliert Kommunikation in einem Netzwerk nicht beliebig. Physikalische Limitierungen, bspw. durch Switches, erschweren es, eine annähernd gleichbleibende Kommunikationsgeschwindigkeit bei steigender Knotenanzahl zu garantieren. Auch die Datenspeicherung unterliegt klaren Einschränkungen. EXASolution, eine massivparallele und spaltenbasierte, relationale in-memory Datenbank kann diese Limitierungen teilweise kompensieren und eine schnelle Analyse großer Datenmengen (100TB bei Clustern von ungefähr 100 Knoten) dennoch ermöglichen. Zudem können Strategien der Datenpartitionierung und -verteilung Zugriffe drastisch beschleunigen. Der Zugriff auf eine große Anzahl ausgelagerter Partitionierungen (Shards) kann durch ein entsprechendes Betriebssystem, bspw. EXAClusterOS, gewährleistet werden, das einen Rechner-Verbund zu einem logischen Knoten zusammenführt [KMN⁺13]. Doch selbst dann wächst die Wahrscheinlichkeit ungünstiger Verteilungen mit der Anzahl der Shards [CAA08, RS10]. Selbstorganisierendes Clustering sowie die Formation hierarchischer Strukturen von ODNs versprechen die Skalierbarkeit weiter drastisch zu erhöhen. Bei einem moderaten System von 5000 Clustern (derartige Infrastrukturen mit ca. einer halben Millionen Servern sind beispielsweise bei Google schon länger im Betrieb), erreicht man bereits den Exabyte Bereich.

3.5 Visual Analytics

Die Transformation von Rohdaten in eine aussagekräftige Visualisierung erfolgt auf Basis einer Modellbildung um wichtige von unwichtigen Rohdaten unterscheiden zu können [KAF⁺08]. Die Modellbildung ist Teil eines Datenabstraktionsprozesses [TVPE92, Tim99], dessen Ergebnis zum Wissenszuwachs in der explorativen Analyse [DMK05] beitragen soll. Sind die ODN Teil eines Geosensornetzwerks [Duc13], so kann die raum-zeitliche Komponente der Daten für weitere Analysen [AA06] sowie zur Qualitätskontrolle genutzt werden. Die interaktive Visualisierung der (Geo)Daten [AA99, WGK10] folgt dem Mantra der Informationssuche: Überblick – Wichtiges herausheben – Zoom und Filter – Details auf Anfrage [CMS99]. In den Schritten *Zoom und Filter* sowie *Details auf Anfrage* ist der semantische Zoom wichtig, der die Präsentation von Datendetails auf Anordnung

vom Nutzer unter Berücksichtigung der Anforderungen bzw. des Ziels der Datenanalyse umgesetzt [Mod97]. Damit ändert sich mit der Zoomstufe nicht nur die Detailliertheit der Information sondern auch die Art bzw. der Typ der dargestellten Information.

3.6 Nutzerinteraktion

Nutzer interagieren mit visuellen Darstellungen auf sehr komplexe Weise – die Qualität der Benutzerschnittstelle hängt häufig davon ab, wie sehr deren Funktionalität mit den Analyseintentionen des Nutzers übereinstimmt bzw. ob die Schnittstelle einzelne Operationen (Toolbox) oder vollständige Analyseabläufe (Processing) unterstützt [BRN12, Tim03]. Untersuchungen zur Funktion von Operationen führten zu einer Unterscheidung in hochwertige (Filter, Relate, Aggregate) und niederwertige (Zoom, Pan) Operationen in der Informationsvisualisierung [Pla05]. MacEachren und Kraak unterscheiden drei verschiedenen Achsen: Interaktionsebene (intensiv bis oberflächlich), Zielgruppe (einzelne Person bis Öffentlichkeit), sowie Aufgabenstellung (Informationsaustausch bis Wissenskonstruktion) [MK01].

4 Benchmarks

Der integrative Charakter von Organic Data bedingt, dass eine Implementierung hinsichtlich verschiedener Kriterien evaluiert werden muss. Im Folgenden wird konkret auf Effizienz, Selbstorganisation, Datensicherheit und Zugänglichkeit eingegangen.

4.1 Benchmark Velocity

Je nach Anwendung müssen Big Data Lösungen unterschiedlichen Ansprüchen bzgl. der Datenverarbeitungsgeschwindigkeit genügen. Offenbar wird dieser Unterschied wenn man bspw. mittels Mobilitätsdaten einen idealen Standort für eine neue Supermarktniederlassung ermitteln oder, im Gegensatz dazu, bei steigender Unfallgefahr den Verkehrsfluss dirigieren möchte. Im Allgemeinen müssen im ODN die Senken bei beliebigen Quelldaten zumindest so schnell bedient werden können, wie für die jeweilige, anwendungstypische Entscheidungsfrequenz nötig. Wir müssen deshalb einerseits zeigen, dass die verwendeten Algorithmen prinzipiell echtzeitfähig sind. Außerdem müssen wir die Auswirkungen lokaler Verarbeitungsalgorithmen und Entscheidungen der ODNs auf das Laufzeitverhalten des Gesamtsystems analysieren. Da sich netzwerkspezifische Kenngrößen, wie z.B. Latenzzeiten, aus einer Vielzahl von Parametern ergeben, bspw. der Netzwerktopologie, dem Datenfluss, sowie lokaler Performanzspezifikationen, werden wir neben ausführlichen Tests anhand der Demonstratoren auf bewährte simulationsbasierte Analyseverfahren zurückgreifen.

4.2 Benchmark Self-X Properties

Es soll ein Design Space für Organic Data Systeme charakterisiert werden, in dem klassische Netzwerkeigenschaften, wie z.B. Bandbreite, Stabilität, den erforschten Netzeigenschaften von Organic Data Networks, z.B. Selbstkonfiguration, -heilung und -optimierung (self-x properties), gegenüber gestellt werden. Eine maßgebliche Frage wird die Abbildung von "Privacy-by-Design" Parametern, also der selbstorganisierte Schutz von Daten, in diesen Designspace haben. Die entwickelten Algorithmen sollen in diesen Designspace eingeordnet und bzgl. ihrer Performanz am Labordemonstrator evaluiert werden. Zukünftigen Nutzern des Systems wird dieser Designspace zusammen mit den entwickelten Algorithmen eine Richtlinie darstellen, anhand derer er sein System auf die konkreten Privacy-Anforderungen und Netzwerkausgestaltung anwenden kann.

4.3 Benchmark Security/Privacy

Um die Sicherheit des resultierenden Organic Data Netzes zu validieren, wird überprüft in wie weit die Aspekte Verschlüsselung, Authentisierung und Anonymität erreicht wurden. Die Verschlüsselung wird dadurch verifiziert, dass die resultierenden Datenströme ununterscheidbar von echtem Zufall sind. Dazu werden unter anderem standardisierte statistische Verfahren zur Überprüfung eingesetzt. Authentisierung wird validiert, indem gezeigt wird, dass ein empfangenes Datenpaket nur von dem erwarteten Sender der Nachricht stammen kann. Der Einsatz moderner Sicherheitsverfahren, wie TLS zur Sicherung der Kommunikation zwischen verschiedenen ODNs sollte dies im allgemeinen gewährleisten. Es ist jedoch im Kontext des Organic Data Netzes erforderlich dies gezielt zu verifizieren. Um Anonymität zu gewährleisten, müssen einzelne Datenquellen voneinander ununterscheidbar sein. Konkret muss gezeigt werden, dass jedes verarbeitete Rohdatenpaket ebenso von mindestens einer weiteren Quelle stammen kann. Damit lässt sich nicht rückverfolgen welche Quelle genau die Rohdaten zur Verfügung gestellt hat: privacy by design.

4.4 Benchmarks Visual Analytics & HCI

Um die Effektivität abstrakter Darstellungsmodi zu analysieren, möchten wir Nutzerbewertungen verschiedener Abstraktionsebenen bei gleichbleibender Analyseaufgabe sammeln. Die Dynamisierung der Visualisierung zur Abbildung sich verändernder Muster und der Adaption der Organic Data Netzwerkstruktur kann durch den Gebrauch von Metriken für Veränderungsrate bemessen werden. Bei der Evaluation der Nutzerinteraktionen als auch deren Umgebungen kommen klassische Methoden wie Zeitmessung, Messung der Anzahl Klicks, Videoanalyse und Befragung zum Einsatz [KEM07]. Die Bewertungskriterien müssen erweitert werden, um die Effektivität horizontaler und vertikaler Darstellungs- und Explorationsmodi zu quantifizieren.

5 Zusammenfassung & Ausblick

In diesem Beitrag haben wir Organic Data als dezentrales Konzept für Big Data dargestellt. Wir haben den Datenfluss in einem Organic Data Netzwerk erläutert und die Spezifikation einzelner Knoten in diesem Netzwerk (Organic Data Nodes) beschrieben. Der notwendigen Integration eines neuen Ansatzes wie Organic Data in den Anwendungskontext haben wir durch Erläuterungen hinsichtlich der Datensicherheit sowie der Usability (Visual Analytics und HCI) Rechnung getragen. Um einerseits die Herausforderungen des Konzepts zu unterstreichen und um andererseits konkrete Implementierungen zu unterstützen, haben wir außerdem konkrete Vorschläge für Benchmark-Metriken unterbreitet. Unsere Ausführungen definieren die Eckpfeiler von Organic Data. Um eine tiefgreifende Integration der verschiedenen Aspekte systematisch umzusetzen, würden wir uns auf den Design Science Ansatz besinnen und schrittweise die Details der verschiedenen verwobenen Entwicklungszyklen herausarbeiten und miteinander in Einklang bringen - von der Anforderungsanalyse zum Feldversuch, vom existierenden Grund- und Expertenwissen zu dessen konkreter Erweiterung und schließlich vom konkreten Systementwurf bis zum marktreifen Prototypen [Hev07].

Literatur

- [AA99] G.L. Andrienko und N.V. Andrienko. Interactive maps for visual data exploration. *International Journal for Geographical Information Science*, 13(4):355–374, 1999.
- [AA06] N. Andrienko und G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data*. Springer Berlin Heidelberg, 2006.
- [BDPA13] Guido Bertoni, Joan Daemen, Michaël Peeters und Gilles Van Assche. The Keccak sponge function family. online: <http://keccak.noekeon.org/>, January 2013.
- [BF10] David Bollier und Charles M Firestone. *The promise and peril of big data*. Aspen Institute, Communications and Society Program Washington, DC, USA, 2010.
- [BR05] John Black und Phillip Rogaway. CBC MACs for Arbitrary-Length Messages: The Three-Key Constructions. *JC*, 18(2):111–131, 2005.
- [BRN12] D. Burkhardt, T. Ruppert und K. Nazemi. Towards process-oriented Information Visualization for supporting users. In *Interactive Collaborative Learning (ICL), 2012 15th International Conference on*, Seiten 1–8, 2012.
- [CAA08] Jean-Daniel Cryans, Alain April und Alain Abran. *Criteria to Compare Cloud Computing with Current Database Technology*, Jgg. 5338, Seiten 114–126. Springer Berlin Heidelberg, 2008.
- [CMS99] S.K. Card, Jock Mackinlay und B. Shneiderman. *Readings in Information Visualization. Using Vision to think*. Morgan Kaufmann Publishers, San Francisco, CA, 1999.
- [CS93] Phillip K Chan und Salvatore J Stolfo. Toward parallel and distributed learning by meta-learning. In *AAAI workshop in Knowledge Discovery in Databases*, Seiten 227–240, 1993.

- [CSF⁺08] D. Cooper, S. Santesson, S. Farrell, S. Boeyen, R. Housley und W. Polk. Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile, May 2008. RFC 5280.
- [DMK05] Jason Dykes, A.M. MacEachren und M.-J. Kraak. *Exploring Geovisualization*. Elsevier Pergamon Press, Amsterdam, 2005.
- [DR08] T. Dierks und E. Rescorla. The Transport Layer Security (TLS) Protocol, Version 1.2, August 2008. RFC 5246.
- [Dre06] Falko Dressler. Self-Organization in Ad Hoc Networks: Overview and Classification. Bericht, University of Erlangen, Dept. of Computer Science 7, Erlangen, 2006.
- [Duc13] Matt Duckham. When Computing Happens Somewhere. In *Decentralized Spatial Computing*, Seiten 3–32. Springer Berlin Heidelberg, 2013.
- [Eas11] D. Eastlake, 3rd. Transport Layer Security (TLS) Extensions: Extension Definitions, January 2011. RFC 6066.
- [fip01] *Federal Information Processing Standards Publication 197 - Announcing the ADVANCED ENCRYPTION STANDARD (AES)*, 26 November 2001. Publication 197.
- [FLMM10] Agostino Forestiero, Emilio Leonardi, Carlo Mastroianni und Michela Meo. Self-Chord: A Bio-Inspired P2P Framework for Self-Organizing Distributed Systems. *IEEE/ACM Transactions on Networking*, 18(5):1651–1664, Oktober 2010.
- [GJH⁺12] C. Grenz, U. Jänen, J. Hähner, C. Kuntzsch, M. Menze, D. d'Angelo, M. Bogen und E. Monari. CamInSens - Demonstration of a Distributed Smart Camera System for In-Situ Threat Detection. In *Distributed Smart Cameras (ICDSC), 2012 Sixth International Conference on*, Seiten 1–2, 2012.
- [GSB02] M. Günes, U. Sorges und I. Bouazizi. ARA - The Ant-Colony Based Routing Algorithm for MANETs. In *Proceedings. International Conference on Parallel Processing Workshops*, Seiten 79–85. IEEE Comput. Soc, 2002.
- [Hev07] Alan R Hevner. A three cycle view of design science research. *Scandinavian journal of information systems*, 19(2):4, 2007.
- [Hol13] Martin Holland. NSA-Überwachungsskandal: PRISM, Tempora und Co. - was bisher geschah. online: <http://heise.de/-1909702>, July 2013.
- [KAF⁺08] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer und Guy Melançon. Visual Analytics: Definition, Process, and Challenges. In Andreas Kerren, John T. Stasko, Jean-Daniel Fekete und Chris North, Hrsg., *Information Visualization*, Jgg. 4950 of *Lecture Notes in Computer Science*, Seiten 154–175. Springer Berlin Heidelberg, 2008.
- [KEM07] A. Kerren, A. Ebert und J. Meyer. *Human-centered Visualization Environments*. Springer Berlin Heidelberg, 2007.
- [KMN⁺13] Alfons Kemper, Tobias Mühlbauer, Thomas Neumann, Angelika Reiser und Wolf Rödiger. Bericht vom Herbsttreffen der GI-Fachgruppe Datenbanksysteme. *Datenbank-Spektrum*, 13(1):65–66, 2013.
- [KS05] S. Kent und K. Seo. Security Architecture for the Internet Protocol, December 2005. RFC 4301.

- [KTGH13] Dominik Klein, Phuoc Tran-Gia und Matthias Hartmann. Big Data. *Informatik-Spektrum*, Seiten 1–5, 2013.
- [MK01] Alan M. MacEachren und Menno-Jan Kraak. Research Challenges in Geovisualization. *Cartography and Geographic Information Science*, 28(1):3–12, 2001.
- [Mod97] D. Modjeska. Navigation in Electronic Worlds: A Research Review, Technical Report. Bericht, Computer Systems Research Group, University of Toronto, 1997.
- [MSSU11] Christian Müller-Schloer, Hartmut Schmeck und Theo Ungerer, Hrsg. *Organic Computing - A Paradigm Shift for Complex Systems*. Autonomic Systems. Birkhäuser Verlag, 2011.
- [MWT11] Hannes Mühleisen, Tilman Walther und Robert Tolksdorf. Data Location Optimization for a Self-Organized Distributed Storage System. In *Proceedings of the Third World Congress on Nature and Biologically Inspired Computing*, Seiten 176–182. IEEE Press, 2011.
- [nis01] NIST Special Publication 800-38B: Recommendation for Block Cipher Modes of Operation; The CMAC Mode for Authentication, 2001.
- [NIS09] NIST. FIPS 186-3: Digital Signature Standard (DSS). Bericht, Information Technology Laboratory, National Institute of Standards and Technology, June 2009.
- [PH94] FJ Provost und DN Hennessy. Distributed machine learning: scaling up with coarse-grained parallelism. In *Proceedings/... International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology*, Jgg. 2, Seite 340, 1994.
- [PKP06] Joel B Predd, SB Kulkarni und H Vincent Poor. Distributed learning in wireless sensor networks. *Signal Processing Magazine, IEEE*, 23(4):56–69, 2006.
- [Pla05] C. Plaisant. Information Visualization and the Challenge of Universal Usability. In J. Dykes, A.M. MacEachren und M.-J. Kraak, Hrsg., *Exploring Geovisualization*, Seiten 53–82. Elsevier Ltd., 2005.
- [RS10] Kurt Rohloff und Richard E. Schantz. High-performance, massively scalable distributed systems using the MapReduce software framework: the SHARD triple-store. In *Programming Support Innovations for Emerging Distributed Applications*, PSI EtA '10, Seiten 4:1–4:5, New York, NY, USA, 2010. ACM.
- [SDv⁺ed] Abbas Sarraf Shirazi, Timothy Davison, Sebastian von Mammen, Jörg Denzinger und Christian Jacob. Adaptive Agent Abstractions to Speed Up Spatial Agent-Based Simulations. *Simulation Modelling Practice and Theory*, 2013 (submitted).
- [SvMJ10] Abbas Sarraf Shirazi, Sebastian von Mammen und Christian Jacob. Adaptive Modularization of the MAPK Signaling Pathway Using the Multiagent Paradigm. In *Parallel Problem Solving from Nature – PPSN XI*, Jgg. 6239 of *Lecture Notes in Computer Science*, Seiten 401–410. Springer Verlag, Krakow, Poland, 2010.
- [SvMJ11] Abbas Sarraf Shirazi, Sebastian von Mammen und Christian Jacob. Hierarchical Self-Organized Learning in Agent-Based Modeling of the MAPK Signaling Pathway. In *CEC 2011, IEEE Congress on Evolutionary Computation*, Seiten 2245–2251, New Orleans, Louisiana, 2011. IEEE Press.
- [SWA08] Padhraic Smyth, Max Welling und Arthur U Asuncion. Asynchronous distributed learning of topic models. In *Advances in Neural Information Processing Systems*, Seiten 81–88, 2008.

- [THH10] Sven Tomforde, Björn Hurling und Jörg Hähner. Dynamic Control of Mobile Ad-hoc Networks-Network Protocol Parameter Adaptation using Organic Network Control. In *ICINCO (1)*, Seiten 28–35, 2010.
- [Tim99] Sabine Timpf. Abstraction, Levels of Detail, and Hierarchies in Map Series. In Christian Freksa und David M. Mark, Hrsg., *Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science*, Jgg. 1661 of *Lecture Notes in Computer Science*, Seiten 125–139. Springer Berlin Heidelberg, 1999.
- [Tim03] S. Timpf. Geographic Activity Models. In Matt Duckham, Michael Goodchild und Michael F. Worboys, Hrsg., *Foundations of Geographic Information Science*, Seiten 241–254. CRC Press, 2003.
- [TVPE92] Sabine Timpf, Gary S. Volta, David W. Pollock und Max J. Egenhofer. A conceptual model of wayfinding using multiple levels of abstraction. In A.U. Frank, I. Campari und U. Formentini, Hrsg., *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, Jgg. 639 of *Lecture Notes in Computer Science*, Seiten 348–367. Springer Berlin Heidelberg, 1992.
- [TZHMS10] Sven Tomforde, Ioannis Zgeras, Jörg Hähner und Christian Müller-Schloer. Adaptive control of sensor networks. In *Autonomic and Trusted Computing*, Seiten 77–91. Springer, 2010.
- [vMS13] Sebastian von Mammen und Jan-Philipp Steghöfer. *The Computer after Me*, Kapitel Bring it on, Complexity! Present and future of self-organising middle-out abstraction. World Scientific Publishing, submitted 2013.
- [vMSDJ11] Sebastian von Mammen, Jan-Philipp Steghöfer, Jörg Denzinger und Christian Jacob. Self-organized Middle-Out Abstraction. In Christian Bettstetter und Carlos Gershenson, Hrsg., *Self-Organizing Systems*, Jgg. 6557 of *Lecture Notes in Computer Science*, Seiten 26–31, Karlsruhe, Germany, 2011. Springer Verlag.
- [vMSSJ12] Sebastian von Mammen, Abbas Sarraf Shirazi, Vladimir Sarpe und Christian Jacob. Optimization of Swarm-based Simulations. *ISRN Artificial Intelligence*, (Article ID 365791):12, 2012.
- [WGH11] Michael Wittke, Carsten Grenz und Jörg Hähner. Towards Organic Active Vision Systems for Visual Surveillance. In Mladen Berekovic, William Fornaciari, Uwe Brinkschulte und Cristina Silvano, Hrsg., *ARCS*, Jgg. 6566 of *Lecture Notes in Computer Science*, Seiten 195–206. Springer, 2011.
- [WGK10] M. Ward, G. Grinstein und D. Keim. *Interactive Data Visualization: Foundations, Techniques, and Application*. A K Peters, 2010.