Zeitschrift: Horizonte : Schweizer Forschungsmagazin
Band: 27 (2015)
Heft: 105

Artikel: 200 Jahre Weltliteratur in 0,4 Sekunden
Autor: Bischofberger, Mirko
DOI: https://doi.org/10.5169/seals-772242

Nutzungsbedingungen
Die ETH-Bibliothek ist die Anbieterin der digitalisierten Zeitschriften. Sie besitzt keine Urheberrechte an den Zeitschriften und ist nicht verantwortlich für deren Inhalte. Die Rechte liegen in der Regel bei den Herausgebern beziehungsweise den externen Rechteinhabern. Siehe Rechtliche Hinweise.

Conditions d'utilisation
L'ETH Library est le fournisseur des revues numérisées. Elle ne détient aucun droit d'auteur sur les revues et n'est pas responsable de leur contenu. En règle générale, les droits sont détenus par les éditeurs ou les détenteurs de droits externes. Voir Informations légales.

Terms of use
The ETH Library is the provider of the digitised journals. It does not own any copyrights to the journals and is not responsible for their content. The rights usually lie with the publishers or the external rights holders. See Legal notice.

Download PDF: 17.11.2024

ETH-Bibliothek Zürich, E-Periodica, https://www.e-periodica.ch

200 Jahre Weltliteratur in 0,4 Sekunden

Langsames Lesen ist out: Heute werden Millionen von Büchern im Nu mit dem Computer durchstöbert und erforscht.
Von Mirko Bischofberger

Alles fing mit Daten an. Mit zu vielen und zu unüberschaubaren Daten, an denen der italienische Jesuit Roberto Busa in den 1940er Jahren arbeitete. Sein Ziel: einen kompletten Index aller 11 Millionen Wörter in den Schriften des Theologen Thomas von Aquin herzustellen. Ein gewaltiges Unternehmen, das ihm wohl in einer Lebenszeit nicht gelingen würde. Doch Vater Busa hatte eine Idee: Eine Maschine solle ihm helfen. Er fand in den Vereinigten Staaten Unterstützung bei IBM-Gründer Thomas Watson. Mit seiner Hilfe war die Indexierung in wenigen Jahrzehnten bewältigt. Der «Index Thomisticus» wurde zu einem wegweisenden, 56-bändigen Werk mit 70 000 Seiten. Es ist das erste Werk, das es dem Benutzer erlaubte, ein komplettes Korpus rasch nach Inhalten zu durchsuchen.

Eine digitale Weltliteratur

Heute dringt die Digitalisierung in alle Bereiche der Geisteswissenschaften vor. «Vor allem die Sprach- und Literaturwissenschaften interessieren sich heute stark für den digitalen Zugang zu ihren Daten», sagt der Zürcher Professor für Computerlinguistik Martin Volk. «Damit können bestimmte Forschungshypothesen mit Zahlen und Statistiken untermauert oder widerlegt werden». In seinem Forschungsprojekt «Text+Berg» hat er alle Bücher des Schweizer Alpen-Clubs digital erschlossen. Das sind 250 Bände, die seit 1864 erscheinen. «Das Material ist eine Fundgrube über die Schweizer Berge», erklärt Volk. «Es zeigt auf, wie sich zum Beispiel das Verständnis der Berge im Laufe der Zeit gewandelt hat. Während die Berge früher als Explorationsobjekte beschrieben wurden, werden sie heute mehr als Sportgeräte angesehen. Das Wort «Wettkampf» kommt heute viel häufiger vor als früher.»

Ein weiteres Schweizer Projekt führt die Universität Genf durch. Dort wird ein Teil der «Bibliotheca Bodmeriana» digitalisiert. Diese aussergewöhnliche Büchersammlung besteht aus über 150 000 Werken in achtzig Sprachen und aus drei Jahrtausenden. Darunter sind das älteste Manuskript des Johannes-Evangeliums aus dem 2. Jahrhundert und die Urschriften der Gebrüder-Grimm-Märchen.

Citizen Science hilft

Doch die Digitalisierung von Büchern ist mühsam. «Die Bücher müssen zuerst von Hand geschnitten und dann alle Seiten separat eingescannt werden», sagt Volk. Er hat im Projekt «Text+Berg» über 120 000 Seiten digitalisiert und weiss, wovon er spricht. «Nach dem Scannen hat man Tausende von Computerbildern und keine Texte.» Für die Texterkennung sind Programme zuständig, die die Buchstaben innerhalb der Bilder erkennen und in Wörter umwandeln. «Die Fehlerrate bei diesem Prozess ist immer noch relativ hoch, gerade bei älteren Schriften aus dem 19. Jahrhundert.» In seinem Projekt ergaben sich etwa zwölf Fehler pro Seite, die man von Hand hätte überprüfen müssen. In der Not wurde Volks Team erfinderisch: Es entwickelte ein Online-Korrektursystem, das es Freiwilligen in einer Art Spiel erlaubte, die Fehler von Hand auszumerzen. Das Citizen-Science-Projekt kam bei den Mitgliedern des Schweizer Alpen-Clubs sehr gut an. «Dank ihrer Hilfe konnten wir in einem halben Jahr über 250 000 Korrekturen durchführen», sagt Volk stolz. Jetzt ist das digitale Bergkorpus praktisch zu 100 Prozent korrekt.

> «Die Google-Applikation Ngram war für die digitalen Geisteswissenschaften wegweisend.»

Sind die Texte digitalisiert, können sie einfach archiviert und angesehen werden. «Vor allem bei alten, raren oder schwierig zugänglichen Dokumenten ist dies sonst nicht möglich», sagt Volk. Das weltweit bekannteste und wohl umfangreichste solche Archiv ist Google Books.

◂ Seite 15 und 16: Beides gibt Aufschluss darüber, mit welchen Lebensmitteln wir kochen: der Blick in den Kühlschrank und das Geschmacksnetzwerk. Die Knotenfarbe steht für die Kategorie, die Grösse für die Häufigkeit in Rezepten. Die Verbindungslinien zeigen, wie viele aromatische Komponenten Lebensmittel teilen.
Bild: Valérie Chételat (S. 15); Yong-Yeol Ahn (S. 16)

Via Volltextsuche können dort die Bestände der Universitätsbibliotheken Harvard, Stanford und New York in Sekundenschnelle durchsucht werden. Auch europäische Bibliotheken, wie diejenige der Universität Oxford oder die Bayerische Staatsbibliothek, sind bereits erfasst.

Freud ist bekannter als Darwin

Auf Google Books aufbauend, entstand 2010 Google Ngram, eine Webapplikation, die das Vorkommen eines Wortes oder einer Wortfolge in allen von Google gescannten Büchern seit 1800 untersucht. Damit können zum Beispiel geschichtliche Ereignisse wie die Abschaffung der Sklaverei untersucht oder auch die sprachliche Veränderung bestimmter Wörter innerhalb einer Sprache beobachtet werden. Sichtbar wird auch die Popularität von Persönlichkeiten im Lauf der Geschichte: Wissenschaftler wie Sigmund Freud, Albert Einstein oder Charles Darwin kommen sehr oft in den Büchern vor, doch Freud wird seit 1950 mindestens doppelt so oft genannt wie die andern.

«Ngram ist bloss ein Beispiel von dem, was heute mit digitalisierten Kulturdaten machbar ist», sagt Jean-Baptiste Michel, Datenwissenschaftler aus Harvard und Autor der Google-Applikation, die bereits Millionen von Benutzern verwenden. Ngram ist aus den digitalen Geisteswissenschaften nicht mehr wegzudenken. Martin Volk von der Universität Zürich bestätigt, dass «Ngram für die digitalen Geisteswissenschaften wegweisend war, vor allem weil es die Methoden in der Breite bekannt gemacht hat».

Über 100 Millionen SMS pro Tag

Die Digitalisierung der bestehenden Literatur ist bloss ein Ansatz der Sprachanalyse. «Heute geben wir über unsere Telefone und Computer pro Tag mehr digitale Texte ein als je zuvor», sagt Elisabeth Stark vom Romanischen Seminar an der Universität Zürich. Allein im Jahr 2013 wurden in Deutschland pro Tag über 100 Millionen SMS versendet. «Alle diese Texte werden zwar so gut wie nie ausgedruckt, sind aber trotzdem Teil unserer Sprachkultur», meint Stark. In ihrem Nationalfonds-Projekt «Sms4science» erforscht sie die sprachlichen Merkmale und die Kommunikation der Schweizer via SMS.

Um an diese Daten zu gelangen, hat Stark zusammen mit andern Forschenden alle Schweizer Handynutzer eingeladen, von SMS eine Kopie an eine Gratisnummer zu schicken. «So konnten wir in der Schweiz rund 26 000 SMS sammeln», sagt Stark. Unter anderem interessiert sie dabei die empirische Erfassung von sprachlichen Ellipsen, also Auslassungen von Wörtern. Typische Beispiele für Ellipsen sind «Komme später» oder «Nervt mich». Um herauszufinden, warum hier das Subjekt fehlt, untersuchte Starks Team alle französischen und deutschen SMS. Dabei stellte es fest, dass die Auslassungen viel seltener sind als angenommen und dass sie denselben Gesetzmässigkeiten wie in der gesprochenen Alltagssprache folgen. «Dies widerspricht dem Eindruck, den man beim Betrachten einzelner SMS häufig hat», sagt Stark, «und deshalb wird eine grosse Datenmenge benötigt, um den tatsächlichen Gegebenheiten in SMS auf die Spur zu kommen.»

Zu wenig Ressourcen in der Schweiz?

Die digitalen Geisteswissenschaften erlauben es, Literatur und Sprache mit Zahlen zu analysieren. Und Zahlen waren schon immer das Markenzeichen der exakten Wissenschaften. Sie ermöglichen es, quantitative Muster und Beziehungen in einer Präzision zu beschreiben, zu der Worte kaum fähig sind. Die nächste Generation von Geisteswissenschaftlern wird demnach datenbasiert arbeiten, so wie es die Bioinformatiker seit Ende des 20. Jahrhunderts tun. «Das Feld wird vor allem durch die enorme Zunahme an digitalen Textmengen angetrieben», sagt Martin Volk. «So wie die Sequenzierung des menschlichen Genoms zur Bioinformatik führte, wird die Digitalisierung unserer Sprache und Literatur unweigerlich in den Geisteswissenschaften bald nicht mehr wegzudenken zu sein.»

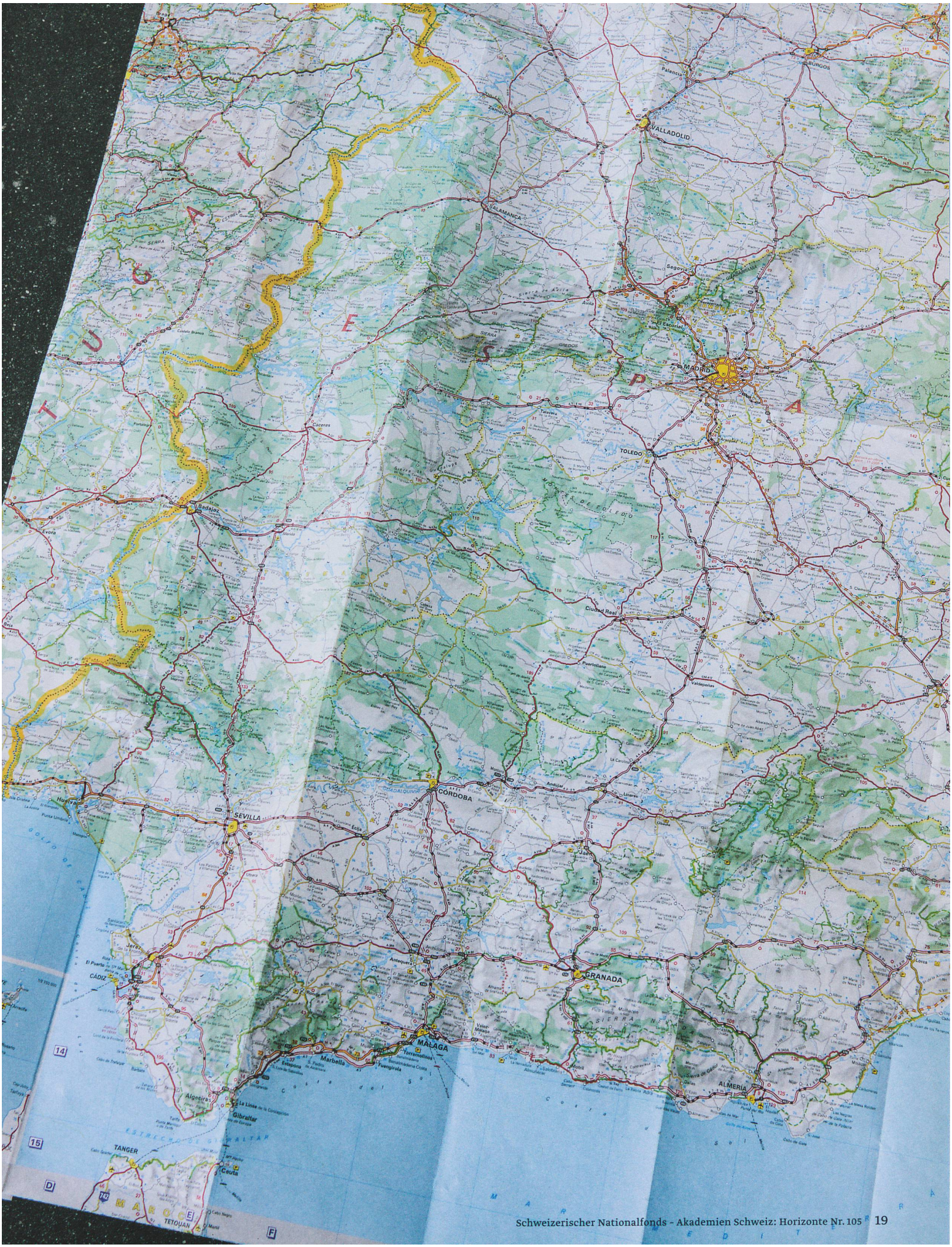
Forschende wie Martin Volk und Elisabeth Stark sind nur der Anfang einer neuen Forschungsära der Geisteswissenschaften. «Leider sind die Ressourcen in der Schweiz für die digitalen Geisteswissenschaften im Moment aber noch gering», sagt Volk. Auch Stark ist dieser Ansicht: «Es gibt an der ganzen Universität Zürich zum Beispiel noch keine Professur für digitale Geisteswissenschaften, obwohl es höchste Zeit dafür wäre». Noch wichtiger scheint beiden der Zugang zu grösseren Datenkonsortien. «Obwohl es wichtige Initiativen auf europäischer Ebene gibt, ist die Schweiz im Moment leider oft nicht Teil davon», sagt Stark. Auch Jean-Baptiste Michel, der bei Google Books ein unglaubliches Reservoir an Daten benutzen durfte, sagt: «Der Zugang zu Daten ist der wichtigste Motor!»

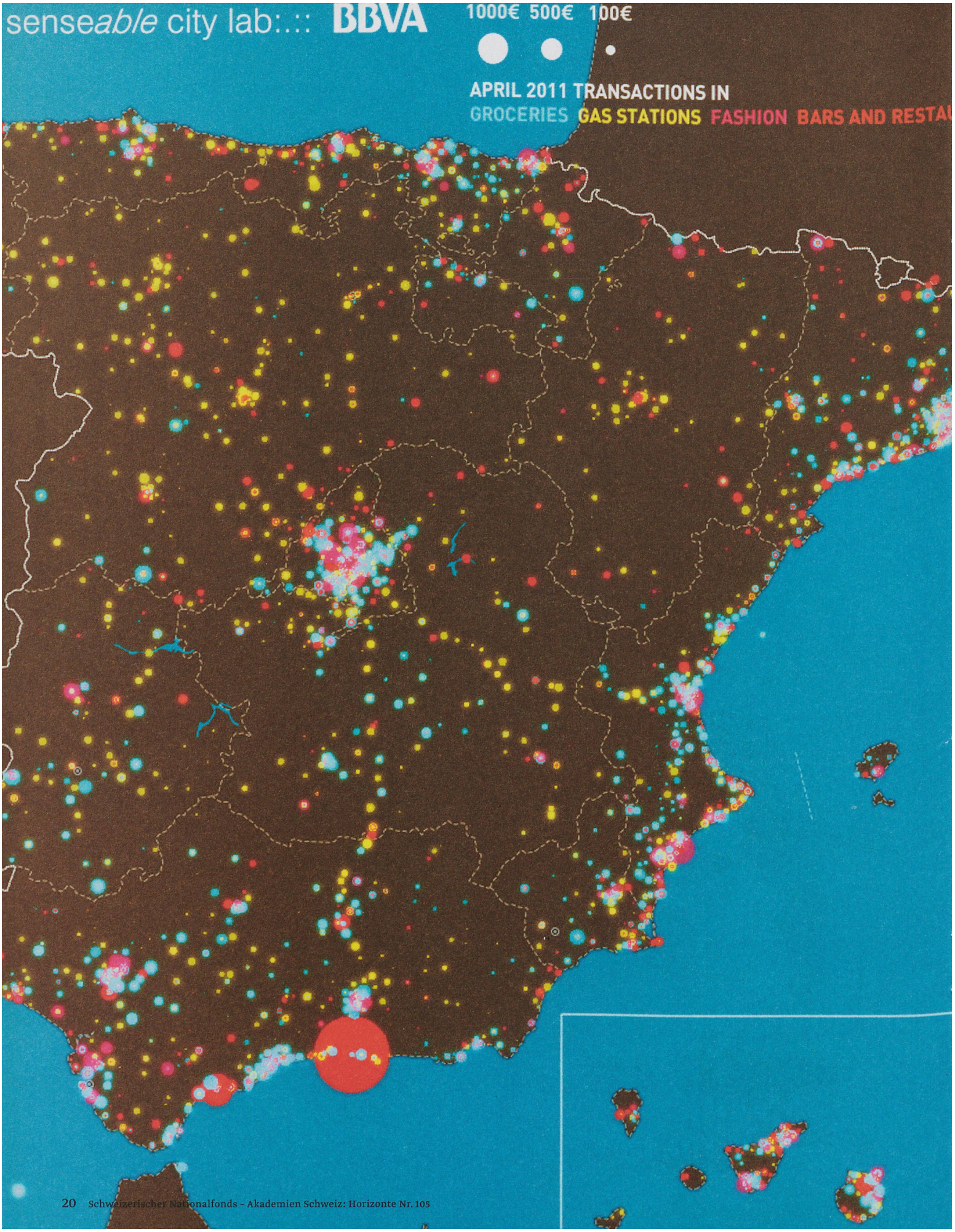
Mirko Bischofberger ist wissenschaftlicher Mitarbeiter des SNF.

—

J.-B. Michel et al.: Quantitative Analysis of Culture Using Millions of Digitized Books. Science. 2011

«Um den tatsächlichen Gegebenheiten in SMS auf die Spur zu kommen, braucht es eine grosse Datenmenge.»


PORTUGAL
ESPAÑA
VALLADOLID
BURGOS
SALAMANCA
Segovia
Avila
El Escorial
MADRID
TOLEDO
Caceres
Badajoz
Evora
Ciudad Real
Puertollano
Valdepeñas
CÓRDOBA
Andújar
Linares
SEVILLA
Huelva
Ecija
Jaen
Jerez
El Puerto
CÁDIZ
Antequera
GRANADA
MÁLAGA
Marbella
Torremolinos
Fuengirola
Estepona
Algeciras
La Línea de la Concepción
Gibraltar
ALMERÍA
TANGER
Ceuta
TETOUAN
MAROC
14
15
D
E
F
Costa del Sol
ESTRECHO DE GIBRALTAR
MARE MEDITERRANEUM


senseable city lab:.:: BBVA
1000€ 500€ 100€
APRIL 2011 TRANSACTIONS IN
GROCERIES GAS STATIONS FASHION BARS AND RESTAU