О. С. Сизов (Компания IDA System)

В 2005 г. окончил Алтайский государственный университет по специальности «эколог-природопользователь». В настоящее время — инженер по обработке ДДЗ компании IDA System (резидент Сколково). С мая 2012 г. (по совместительству) — научный сотрудник Института криосферы Земли СО РАН. Кандидат географических наук.

Вопросы практического внедрения Big Geo Data на примере развития технологий дистанционного зондирования

КРАТКАЯ ИСТОРИЯ И ОПРЕДЕЛЕНИЕ ТЕРМИНОВ BIG DATA И BIG GEO DATA

Проблемы практического использования большого количества данных связаны с экспоненциальным ростом объема информации в обществе, получившим название «информационный взрыв» (Information Explosion). Наглядно этот феномен прослеживается на примере роста количества печатных книг в Европе с начала изобретения книгопечатания в середине XV в. (рис. 1).

Сам термин «информационный взрыв» впервые упоминается в 1941 г. в словарной статье Oxford English Dictionary [2], а уже в 1944 г. библиотекарь Уэслианского университета (Коннектикут, США) Фремонт Райдер (Fremont Rider) публикует работу, в которой приходит к выводу, что число книг в библиотеках американских университетов удваивается раз в 16 лет [3].

Новый этап информационного развития связан с изобретением ЭВМ, повсеместным переходом к цифровым способам хранения и передачи данных, а также с появлением сети Интернет (рис. 2). По оценке компании Cisco, в ближайшее время мы вступим в эру зетабайтов (The Zettabyte Era) [4], т. е. к концу 2016 г. ежегодный IP-трафик преодолеет порог в 1000 экзабайт и будет увеличиваться на 2 зетабайта в год вплоть до 2019 г. (рис. 3).

Термин «большие данные» (big data, BD) в научный оборот впервые был введен

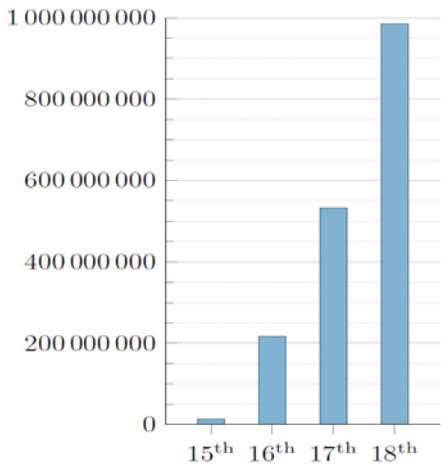

Рис. 1. Число печатных книг в Европе (без Турции и России) за 1450–1800 гг., шт. [1]

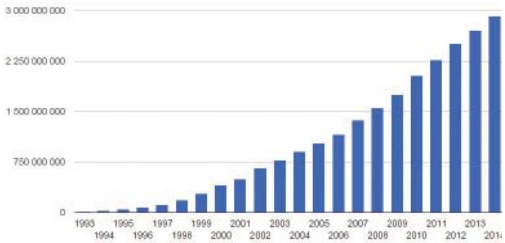

Рис. 2. Число пользователей сети Интернет, чел. [5]

Рис. 3. Прогноз роста интернет-трафика по версии компании Cisco, Tbps [4]

в 1997 г. инженерами Intel и NASA Майклом Коксом (Michael Cox) и Дэвидом Элсворфом (David Ellsworth), изучавшими проблему визуализации больших массивов информации [6]. А в 2001 г. аналитик компании Meta Group (поглощена Gartner) Дуглас Лэйни (Doug Laney) представил работу «3D Data Management: Controlling Data Volume, Velocity and Variety» [7], где предложил основные характеристики, которым должны отвечать большие данные, так называемые «3V»:

1. Объем (англ. volume, величина физического объема).

2. Скорость (англ. velocity, как скорость прироста, так и необходимость высокоскоростной обработки и получения результатов).

3. Многообразие (англ. variety, возможность одновременной обработки различных типов структурированных и полуструктурированных данных).

Часто к трем «V» добавляют еще одну — Veracity (достоверность), понимая под этим целостность данных, их способность к структурированию и уровень доверия к результатам [8].

В настоящее время основной поток геопространственной информации, отвечающей общим критериям Big Data, генерируется с помощью:

- глобальных систем позиционирования;
- аппаратуры дистанционного зондирования Земли, установленной на БПЛА, самолетах и космических спутниках;
- глобальных систем позиционирования (GPS, ГЛОНАСС, Beidou и др.);
- локальных сенсоров, привязанных к определенному объекту или точке с известными координатами (датчики уровней воды на реках, логгеры метеопараметров, различные датчики мобильных устройств и др.);
- меток радиочастотной идентификации (RFID), установленных на подвижных объектах;
- социальных сетей с географической привязкой контента (Twitter, Facebook и др.) [9, 10].

Кроме этого, к большим данным, обладающим пространственной компонентой, или, для краткости, большим геоданным (Big Geo Data), можно отнести весь накопленный архив географических знаний (научные труды, картографические произведения, результаты натурных наблюдений), которые становятся доступны для анализа в цифровом виде.

Таким образом, лавинообразное накопление больших массивов данных и интенсификация информационного обмена приводят к возникновению ряда технологических и методологических проблем, решение которых можно объединить в рамках единого направления Data-driven Geography — географических исследований, определяемых данными [11].

ОРГАНИЗАЦИЯ РАБОТЫ С BIG GEO DATA НА ПРИМЕРЕ EOSDIS

Поскольку осознание проблем в области управления, анализа, хранения и распространения больших объемов информации происходило эмпирически по мере развития технических возможностей, наглядно процесс развития и усложнения информационной инфраструктуры можно показать на примере практической реализации программы космического мониторинга Земли EOS (NASA's Earth Observing System).

Реализацию программы EOS обеспечивает система EOS DIS (EOS Data and Information System), главная цель которой состоит в том, чтобы обеспечить сбор, хранение, обработку данных мониторинга Земли и справочной информации, предоставление удаленного доступа к каталогам, а также распределение данных исходя из потребностей максимального числа пользователей [12].

В настоящее время источником первичной информации являются 22 активные миссии [13], осуществляющие мониторинг всех компонентов природной среды в глобальном масштабе и режиме, приближенном к реальному времени (рис. 4).

По состоянию на сентябрь 2012 г. общий объем накопленных в системе данных составил примерно 10 петабайт. Суммарный объем поступающих с целевой аппаратуры космических аппаратов исходных («сырых») данных и продуктов стандартных уровней обработки превышает 8,5 терабайт в день. Система обеспечивает онлайн-доступ к более чем 2800 наборам данных и сервисам для более чем 185 000 пользователей на бесплатной основе [14].

Рис. 4. Временная шкала жизненного цикла научных миссий EOS с 2001 г. (включая планируемые проекты) [14]

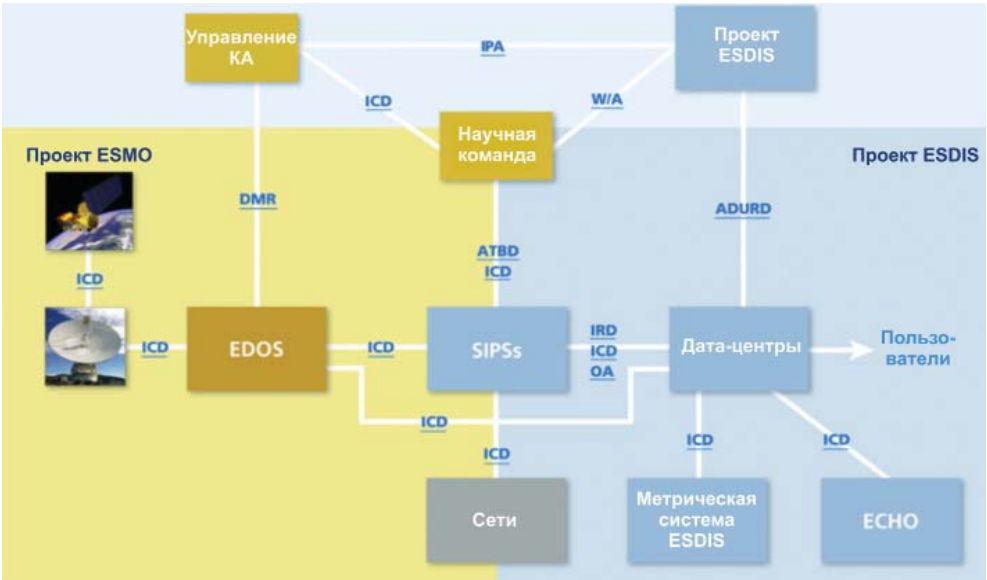

Рис. 5. Функциональная структура EOSDIS [12]

Вычислительные потребности системы с 2008 г. обеспечивает суперкомпьютер Pleiades, установленный в NASA Advanced Supercomputing (NAS) исследовательского центра Эймса (NASA Ames). Суперкомпьютер Pleiades при пиковой нагрузке процессоров (11 312 узлов, 211 360 ядер) — большие проблемы обеспечивает скорость вычислений 5.34 петафлопс, объем оперативной памяти составляет 724 Tb, объем хранилища данных — 1,3 петабайт [15].

Основная цель EOSDIS изначально заключалась в создании системы с распределенной открытой архитектурой. EOSDIS версии "0" появилась в августе 1994 г. как интегратор систем, которые имели функциональные возможности управления информацией в семи распределенных центрах-архивах DAAC (Distributed Active Archive Centers) и были связаны через внутреннюю сеть для взаимодействия друг с другом [16].

В настоящее время в EOSDIS применяется распределенная архитектура CINTEX (Catalog Interoperability Experiment), которая позволяет располагать элементы системы в различных местах, чтобы в максимальной степени использовать различные научные возможности. Для природно-ресурсных спутников EOSDIS предоставляет возможности по управлению, контролю, планированию, сбору данных и начальной обработке (уровень 0) (проект ESMO — Earth Science Mission Operations). Последующие усилия направлены на решение задач тематической обработки данных в рамках проекта Earth Science Data and Information System (ESDIS). Технология обработки включают: создание научных продуктов более высокого уровня (уровень 1–4) для миссий EOS; архивирование и распространение продуктов миссий EOS и других спутников, а также данных аэросъемки и полевых измерений. Обобщенная функциональная структура системы EOSDIS представлена на рис. 5.

Стандартные продукты уровней 1–4 представляют собой результаты глубокой

предварительной или тематической обработки, подготовленные исходя из потребностей научного сообщества, распределенные в пространстве и/или во времени и доступные, если имеются первичные данные для их создания [12]. В этом отношении ключевая особенность архитектуры EOSDIS состоит в том, что для решения задач по научному обоснованию и программной реализации процедур формирования итоговых информационных продуктов привлекаются внешние научные команды на конкурсной основе. Подобная форма взаимодействия NASA с мировым научным сообществом получила название Science Investigator-led Processing Systems (SIPS). Такой подход позволяет использовать наиболее актуальные и производительные алгоритмы обработки данных и обеспечивать максимальное качество производных продуктов.

Все научные операции SIPS выполняются в рамках распределенной системы нескольких взаимосвязанных узлов с конкретными обязанностями для создания, архивирования и распределения научных данных о Земле. После этого созданные в SIPS продукты направляются в соответствующие DAACs для архивации и распространения [16]. Распределенные центры обработки данных обслуживают сообщество пользователей путем предоставления возможности для поиска, визуализации данных и прямого доступа к данным и специализированным услугам.

Таким образом, в процессе эволюции системы EOSDIS в NASA уже в 2007 г. были сформированы основные потребности пользователей на новом этапе развития научных данных и вычислительной среды, которые заключаются в следующем:

- обеспечение всеобщей устойчивой инфраструктуры данных;
- обеспечение максимальных вычислительных возможностей;
- обеспечение ресурсов для эффективного моделирования природных процессов и явлений;
- интенсификация продуктивности научных исследований (методов, алгоритмов и инструментов) как процесса получения новых знаний [14].

Для реализации возникших потребностей в настоящее время развивается система NASA Earth Exchange (NEX) (рис. 6), которая, по сути, является единой платформой для исследований в области наук о Земле, обеспечивающей:

- вычислительные возможности передового суперкомпьютера Pleiades;
- возможности реализации алгоритмов в привычной среде IDL, MatLab, R, Python и др.;
- возможности моделирования природных компонентов, включая предоставление доступа к существующим моделям;
- весь архив дистанционных данных, накопленный в рамках EOS;
- рабочую среду для управления проектом, в которой пользователь может удаленно использовать все технические возможности;
- инфраструктуру для распространения полученных результатов, сравнения данных, участия в новых или существующих проектах и др. [17].

ОСНОВНЫЕ ПРОБЛЕМЫ И ПОДХОДЫ ДЛЯ НАУЧНО-ПРАКТИЧЕСКОГО ПРИМЕНЕНИЯ BIG GEO DATA

На примере эволюции системы EOSDIS хорошо видно, как по мере накопления все больших массивов данных происходит адаптация всей информационной инфраструктуры к новым условиям, которые диктуют необходимость постоянного увеличения вычислительных мощностей, перехода к распределенной структуре хранения и обработки, совершенствования измерительной аппаратуры и развития методологического аппарата научных исследований.

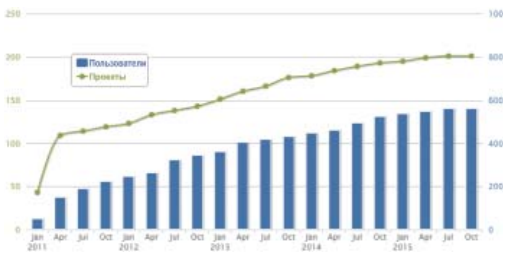

Рис. 6. Динамика числа пользователей в системе NASA Earth Exchange (NEX) [18]

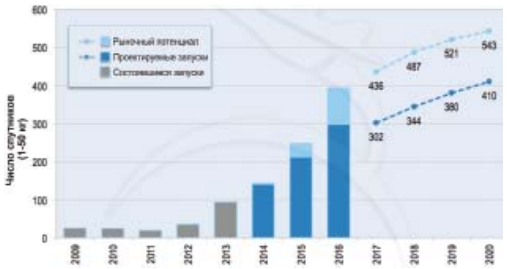

Рис. 7. Фактический и прогнозируемый рост числа микро- и наноспутников на орбите [19]

И если раньше, в условиях централизации источников первичных наблюдений, подобные проблемы возникали только применительно к узким отраслевым направлениям, которые решались силами крупных государственных акторов (NASA), то в настоящее время можно говорить о широкомасштабном развитии распределенной сети наблюдений, продуцирующей избыточный информационный поток.

Ярким примером децентрализации наблюдений является существенное увеличение на орбите микро и мини спутников (менее 50 кг), значительная часть которых предназначена для мониторинга Земли (рис. 7). При этом стоимость создания, запуска и эксплуатации космических аппаратов теперь могут себе позволить не только небольшие страны, но и высокотехнологичные частные компании (Google/Skybox Imaging, Planet Labs, OmniEarth и др.).

Не менее революционные изменения происходят и в сегменте беспилотных летательных аппаратов, где разнообразие платформ для индивидуального пользования наряду с широким выбором компактных мульти- и гиперспектральных сенсоров массой не более 2 кг (Airinov, Micasense, Tetracam, Rikola, HeadWall, Gamaya и др.) создает возможности независимого, оперативного и относительно дешевого сбора высокодетальных пространственных данных для любого заинтересованного пользователя, не обязательно обладающего профессиональными навыками.

При этом при использовании больших объемов различных дистанционных данных возникают методологические проблемы, которые имеют общие черты с вызовами, установленными для Big Data [20]. В целом их можно разделить на три группы:

1. «Количество не значит полнота». Проблемы данной группы связаны с техническими особенностями автоматического сбора информации, при котором далеко не всегда удается обеспечить равные объемы наблюдений для различных классов объектов или, в географии, для различных участков земной поверхности. Это означает, что видимое количество доступных космических снимков отнюдь не гарантирует того, что конкретный район будет обеспечен съемкой в нужный период или с заданным временным интервалом. Пространственную неравномерность сложно учитывать при проведении конкретных исследований, поскольку обычно приоритет отдается задаче, т. е. сначала происходит постановка задачи, а затем уже идет сбор доступной информации. Поэтому в исследованиях, определяемых данными (Data-driven Geography), перечень доступных для решения задач может существенно отличаться от перечня практически значимых задач.

2. «Количество не значит чистота». Для big data преобладание и даже доминирование несортированных данных над сортированными является непременным условием. При дистанционных исследованиях на стадии проектирования сенсора есть возможность минимизировать количество шума, устранить механические погрешности аппаратуры, провести калибровку. Но даже при корректной работе аппаратуры существует большое количество внешних воздействий, которые делают полученную информацию непригодной для практического использования. Оценка пригодности данных и удаление информационного шума не могут опираться исключительно на автоматизированные алгоритмы, за исключением очевидных случаев (наличие облачности, снежного покрова). Поэтому для большинства задач должен формироваться индивидуальный набор критериев проверки качества с учетом базы знаний в каждой тематической области (учет фаз вегетации растительности, гидрологического цикла водных объектов и т. п.). Такая проверка, в частности, может строиться на основе сочетания логических и семантических правил [11].

3. «Корреляция не значит причина». Статистические методы анализа больших массивов данных могут ошибочно приводить к выявлению корреляции логически не связанных явлений. В этом отношении часто встречаются примеры взаимосвязи между глобальным потеплением и сокращением числа пиратов (рис. 8) [21] или развития депрессии у людей, укушенных кошками [22]. Автоматическая классификация космических снимков методами без обучения также зачастую приводит к объединению пикселей, имеющих сходные значения яркости, но логически относящихся к различным типам земной поверхности. При этом полученные результаты могут найти лишь ограниченное практическое применение в силу слабой сопоставимости

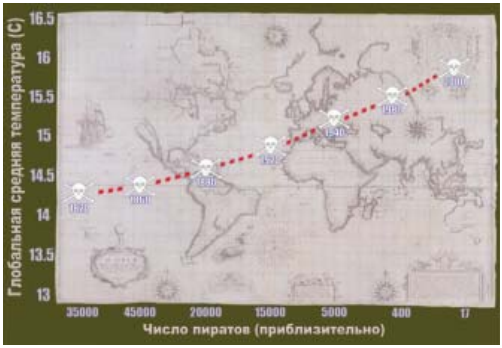

Рис. 8. Зависимость глобального потепления от числа пиратов [21]

даже с результатами, полученными на основе снимка за другую дату или снимка с отличающимися техническими параметрами.

Тем не менее научно-практическое внедрение методов Data-driven Geography, по всей видимости, вопрос ближайшего будущего. Их сложно рассматривать как альтернативу традиционным методам получения новых знаний, скорее они представляют интерес при решении таких задач, как:

- сбор и первичный анализ данных, постановка рабочей гипотезы;
- визуализация различных типов пространственных данных;
- моделирование процессов с учетом теоретических и логических построений [9].

С этой точки зрения можно предложить общие подходы, которые могут облегчить процесс работы с Big Geo Data в рамках решения определенной исследовательской задачи:

1. Стандартизация измерительной аппаратуры – применительно к любым типам сенсоров и датчиков (включая БПЛА, наземные приборы и логгеры, устанавливаемые in situ) должны применяться процедуры, подтверждающие заявленную точность и корректность получаемых первичных

данных. Это позволит существенно сократить усилия по структурированию, валидации и фильтрации.

2. Стандартизация форматов и метаданных – даже для небольших единичных исследований и проектов необходимо предусматривать возможности хранения и передачи итоговой информации на основе наиболее распространенных стандартов, что сократит в будущем усилия по унификации базы знаний для решения аналогичных задач.

3. Историчность исследований – выполнение необходимых действий по сбору и переводу в цифровой вид любой архивной информации (базы знаний), которая была накоплена по заданному объекту. Ретроспективный анализ и сочетание результатов, полученных на основе различных методов, позволят выявить ошибки в обоих случаях.

4. Алгоритмизация — представление накопленной базы знаний по каждой отраслевой задаче (знания специалистов, логические правила, результаты ранее выполненных работ и др.) в унифицированном виде (выделение обязательных технологических блоков обработки данных, определение критериев самопроверки и др.).

5. Сходимость и воспроизводимость результатов — использование различных исходных данных, моделей и алгоритмов при автоматизированном решении конкретной задачи, что позволит оценить достоверность полученных результатов, определить процент погрешности и воспроизвести результат в будущем при необходимости.

Таким образом, несмотря на наличие отдельных методологических вопросов, существуют подходы, которые позволяют реализовать новые возможности обработки больших пространственных данных, что в конечном счете выводит географические исследования на более высокий уровень точности и производительности.

ПЕРСПЕКТИВЫ ИССЛЕДОВАНИЙ НА ОСНОВЕ BIG GEO DATA

Новый уровень географических исследований в методологическом плане связан в первую очередь с интеграцией и синергией абдуктивных, индуктивных и дедуктивных подходов при решении тематических задач и получении новых предметных знаний [23].

В методическом отношении использование методов Data-driven Geography позволяет говорить о переходе к «точной географии», где для каждого участка земной поверхности с заданной размерностью и погрешностью определения положения могут быть сформированы наборы данных по каждому компоненту природной среды на определенный момент времени. При этом по мере накопления данных будут формироваться временные ряды, на основе которых появятся возможности детального моделирования динамики геопроцессов.

Технологический и информационный прорыв, который привел к децентрализации и увеличению количества источников первичной географической информации, в перспективе существенно ускорит интеграционные процессы, в том числе в вопросе обмена результатами исследований. В частности, это соответствует одному из принципов крупнейшей сети персональных метеостанций Weather Underground — «вместе мы знаем больше (Together We Know More)» [24].

Можно также отметить, что с развитием «персонализации» географических исследований формируется большой круг задач, в которых методы Data-driven Geography могут получить максимальное применение. Среди таких задач — навигация (сервисы общедоступных треков), создание открытых карт (проект OpenStreetMap), мониторинг экстренных ситуаций (краудсорсинг при поиске пропавших летательных аппаратов),

Рис. 9. Структурная схема организации научной деятельности в системе EOS [14]

идентификация объектов (проект Wikimapia, сервисы геопозиционированных фотографий, панорамы улиц) и др.

Таким образом, анализ литературы [14, 25] и результаты проведенного исследования позволяют выделить основные условия перехода к научно-практической работе с Big Geo Data, в числе которых:

1. Оборудование (технологии): обеспечение максимальной вычислительной мощности и скорости выполнения алгоритмов сбора, анализа, передачи и хранения больших объемов структурированных и плохо структурированных данных.

2. Методология (исследования): при выявлении закономерностей —создание аналитического инструментария с опорой на методологический аппарат, разработанный и апробированный для каждой сферы деятельности.

3. Задачи (приложения): постановка ясных задач для машинного анализа с привлечением отраслевых специалистов, обеспечивающих корректное формирование обучающей выборки данных и выявление реальных взаимосвязей.

4. Образование (понимание): модернизация образовательных программ исходя из современных технологических возможностей анализа данных, что позволит адекватно подготовить будущих специалистов для работы в новой информационной парадигме.

Визуально сочетание описанных выше условий можно представить в виде общей схемы организации исследований, предложенной в рамках развития системы EOS (рис. 9).

СПИСОК ЛИТЕРАТУРЫ

1. Eltjo, B., Luiten, J.Z. (2009). Charting the «Rise of the West»: Manuscripts and Printed Books in Europe, A Long-Term Perspective from

the Sixth through Eighteenth Centuries», The Journal of Economic History, Vol. 69, No. 2, 409-445.

2. http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

3. Rider, F. (1944). The Scholar and the Future of the Research Library. New York City: Hadham Press.

4. The Zettabyte Era: Trends and Analysis / Cisco Visual Networking Index (VNI): Forecast and Methodology, 2014–2019. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.pdf

5. http://www.internetlivestats.com/internet-users/

6. Cox, M., Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. In Proceedings of the 8th conference on Visualization '97 (VIS '97). IEEE Computer Society Press, Los Alamitos, CA, USA, 235-ff.. https://www.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf

7. Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety, Technical report, META Group. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

8. «What is Big Data?». Villanova University. http://www.villanovau.com/resources/bi/what-is-big-data/#.VfLqMBHtlBc

9. Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging? Journal of Regional Science, 50(1), 181–201

10. Sui, D., Goodchild, M. F. (2011). The convergence of GIS and social media: Challenges for GIScience. International Journal of Geographical Information Science, 25(11), 1737–1748.

11. Miller H., Goodchild M. (2015). Data-driven geography. GeoJournal, Vol. 80, Iss. 4, 449-461. DOI: 10.1007/s10708-014-9602-6

12. https://earthdata.nasa.gov/

13. http://eospso.gsfc.nasa.gov/content/all-missions

14. Science Plan for NASA's SMD 2007-2016. NASA Headquarters, 2007. 170. http://science.nasa.gov/media/medialibrary/2010/03/31/Science_Plan_07.pdf

15. http://www.nas.nasa.gov/hecc/support/kb/entry/77

16. Maiden M. (2011). Philosophy and Architecture of the EOS Data and Information System // Remote Sensing and Digital Image Processing. Springer New York. 35-47. http://dx.doi.org/10.1007/978-1-4419-6749-7_2

17. NASA 2014 Science Plan. NASA Headquarters, 2014. 123. http://science.nasa.gov/media/medialibrary/2015/06/29/2014_Science_Plan_PDF_Update_508_TAGGED.pdf

18. https://nex.nasa.gov/nex

19. Low Earth Orbit Satellite Volume Set to Rapidly Expand. Earth Imaging Journal. 8 October 2014. http://eijournal.com/news/industry-insights-trends/low-earth-orbit-satellite-volume-set-to-rapidly-expand

20. Mayer-Schonberger, V., Cukier, K. (2013). Big Data: A revolution that will transform how we live, work, and think.

21. Henderson, B. (2005). Open Letter To Kansas School Board. Venganza.org. http://www.venganza.org/about/open-letter/

22. Hanauer, D., Ramakrishnan, N., Seyfried, L. Describing the Relationship between Cat Bites and Human Depression Using Data from an Electronic Health Record, PLoS ONE, vol. 8, no. 8, 2013, e70585.

23. Gahegan, M. (2009). Visual exploration and explanation in geography: Analysis with light. In H. J. Miller & J. Han (Eds.), Geographic data mining and knowledge discovery (2nd ed., pp. 291-324). London: Taylor and Francis

24. http://www.wunderground.com/about/background.asp

25. Boyd, D., Crawford, K. (2012). Critical Questions for Big Data. Information, Communication & Society. Vol. 15, Iss. 5, 2012. DOI:10.1080/1369118X.2012.678878