MAPPING netCDF DATA TO RELATIONAL MODEL TO ENABLE ANALYTIC PROCESSING OF LARGE COLLECTIONS OF REMOTE SENSING DATA

Dmitri L. Chubarov[1], *Nikolay N. Dobretsov*[1,2], *Vladimir A. Kikhtenko*[1]

[1] Institute of Computational Technologies SB RAS, Novosibirsk, Russia
[2] V.S. Sobolev Institute of Geology and Mineralogy SB RAS, Novosibirsk, Russia

Abstract

We present a relational schema for the data and metadata contained in MODIS products represented as a collection of HDF files. Commonly used approach formalized in the NetCDF model is based on storing the data and metadata of one scene in a single file. However to develop tools that support research that requires the analysis of time series of remote sensing data new data models may be needed. One promising approach is based on the indexing and addressing the data at the level of individual pixels. The relational schema presented in this paper is intended towards the representation of the metadata present in NetCDF model during the transition towards indexing of individual pixels.

Keywords: databases, remote sensing, time series, research software, metadata.

ОТОБРАЖЕНИЕ МОДЕЛИ ДАННЫХ NETCDF В РЕЛЯЦИОННУЮ МОДЕЛЬ ДЛЯ РАБОТЫ С КОЛЛЕКЦИЯМИ ДАННЫХ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ

Чубаров Д.Л.[(1)], Добрецов Н.Н.[(1)(2)], Кихтенко В.А.[(1)]

[(1)] Институт вычислительных технологий СО РАН, г. Новосибирск

[(2)] Институт геологии и минералогии им. В.С. Соболева СО РАН, г. Новосибирск

В работе предложена реляционная схема для данных и метаданных информационных продуктов MODIS, представленных в виде множества файлов в формате HDF. Существующий подход к хранению данных дистанционного зондирования, формализованный в модели netCDF, основан на хранения всех данных, объединенных одной сценой, и сопутствующих им метаданных в одном файле. Для разработки программных инструментов, предназначенных для поддержки решения задач анализа временных рядов измерений, полученных по данным дистанционного зондирования, требуются новые модели данных, одной из которых является индексирование данных на уровне отдельных пикселей. Предложенная в работе схема, позволяет представить сопутствующие метаданные при переходе от схемы netCDF к индексированию отдельных пикселей.

Ключевые слова: базы данных, данные дистанционного зондирования, временные ряды, информационные системы поддержки научных исследований, метаданные.

Введение. Одним из аспектов роста объема доступных данных дистанционного зондирования, является появление возможности исследования временных рядов измерений [1]. Для обеспечения возможности работы с временными рядами измерений вместо отдельных сцен, характеризирующихся единым интервалом измерений для всех составляющих сцену пикселей, представляет интерес переход к возможности использования и индексирования архива данных дистанционного зондирования на уровне отдельных пикселей [2]. Такой переход ставит вопрос о том, какие изменения необходимы в представлении сопутствующих метаданных.

С середины 1980-х годов при работе с данными дистанционного зондирования и в целом, с данными измерений, преобладал подход, когда данные и метаданные хранятся совместно в блоках, размер которых определяется возможностями эффективной передачи информации, т.е. емкостью используемых переносных носителей или пропускной способностью сетей связи. При этом каждый блок содержит сопроводительную информацию, необходимую для обработки содержащихся в нём данных, начиная от способа представления числовых величин в виде битовой последовательности и заканчивая инструкциями по интерпретации извлекаемых значений.

Для обеспечения хранения совместного блокового хранения данных и метаданных был разработан единый формат Common Data Format (CDF). На основе идей, заложенных в CDF [3], были независимо разработаны форматы NetCDF [4] и HDF [5], которые используются и эволюционируют по сей день, а их дальнейшее развитие согласовано таким образом, что формат NetCDF4 может рассматриваться как модель представления данных, основанная на HDF5 [6]. На основе принципов, заложенных при создании этих форматов, реализован протокол DAP [7], ориентированный на передачу данных и метаданных в распределенных информационных системах.

Развитие технологий хранения больших объемов данных, а также увеличение пропускной способности сетей связи и резкий скачок в повышении производительности вычислений, привели к осознанной потребности в предоставлении доступа ко всему объему имеющихся данных. Представление всего объема данных без деления на блоки ориентировано на архитектуру, в которой все данные находятся постоянно в онлайн доступе, а вычисления над ними выполняются по запросу. Один из способов реализации этого принципа состоит в индексировании отдельных пикселей, когда каждый пиксель каждого спутникового снимка, представленного в архиве, может быть непосредственно поименован. Это не единственный возможный

способ. Другой способ состоит в расширении языка запросов функциями для выполнения операций с массивами пикселей. При этом результатом выполнения запроса является также массив пикселей или последовательность таких массивов [8–10].

Реализация любого из перечисленных выше механизмов сдерживается, в том числе, трудностями переноса метаданных в формируемые структуры данных без деления на блоки. Кроме того, при формировании временных рядов измерений возникают новые метаданные и/или вспомогательные данные, описывающие специфические свойства и характеристики, присущие уже временным рядам, и необходимые для их последующей обработки.

Метаданные для данных дистанционного зондирования. Основные структуры данных, которые используются для хранения результатов измерений, это скаляры, векторы и многомерные массивы. Для представления измерений, связанных с объектами на поверхности Земли, выделяют также точечные измерения, измерения, заданные на регулярных сетках, также называемые растрами, и измерения, выполненные при последовательном сканировании [11].

Метаданные, сопровождающие данные дистанционного зондирования, также можно разделить на несколько уровней в зависимости от того, относятся ли они к отдельной сцене, к ряду сцен, содержащих данные одного типа – информационному продукту, или к отдельному пикселю (таблица 1).

Сами метаданные подразделяются на те, что связаны с форматом хранения данных и применяются при преобразовании из одного формата хранения в другой, а также *тематические* метаданные, которые необходимы для дальнейшей обработки (таблица 2).

Метаданные уровня пикселей могут быть представлены в той же структуре, что и сопровождаемые ими измерения. Для хранения метаданных продукта и сцены необходимы свои структуры данных. В случае представления данных в реляционной схеме этими структурами могут быть три различные таблицы, о которых пойдет речь в следующем разделе (рис. 2).

Таблица 1. Уровни метаданных.

Уровень метаданных	**Примеры**
сцена	время съёмки, изображенная область поверхности Земли, число пикселей
информационный продукт	название продукта, название прибора, описание значений битового поля оценки качества, типы данных, допустимые значения данных
пиксель	зенитный угол обзора, зенитный угол солнца, координата центра, изображенная область поверхности Земли, битовое поле оценки качества

Таблица 2. набор тематических метаданных, сопровождающих коэффициенты спектральной яркости, восстановленные по данным MODIS (MOD09) [12].

Наименование поля	**Уровень**
Число измерений, использованных для восстановления значения в разрешении 1км	пиксель
Состояние по завершении алгоритма обработки	пиксель
Описание значений поля состояния	продукт
Зенитный угол сенсора	пиксель
Азимутальный угол сенсора	пиксель
Расстояние до сенсора	пиксель
Зенитный угол Солнца	пиксель
Азимутальный угол Солнца	пиксель
Состояние по завершении алгоритма геопривязки	пиксель
Число измерений, использованных для восстановления значения в разрешении 500м	пиксель
Код контроля качества данных для разрешения 500м	пиксель
Описание значений кода контроля качества	продукт
Код контроля качества данных для разрешения 250м	пиксель
Описание значений кода контроля качества	продукт
Площадь пересечения изображения пикселя, зарегистрированного сенсором, и ячейки сетки	пиксель
Состояние по завершении алгоритма обработки	пиксель

Реляционная схема для данных и метаданных на примере. Для того чтобы проиллюстрировать важность сохранения всех уровней метаданных при работе с временными рядами измерений, рассмотрим следующий пример, связанный с исследованием в области применения технологий спутникового мониторинга в сельском хозяйстве.

Задача: *Для выполнения исследования состояния посевов необходимо для заданного поля пшеницы, представляющего собой однородный участок поверхности Земли, выбрать все пиксели, в которых поверхность Земли не была закрыта облачностью, а уровень содержания аэрозольных частиц в атмосфере был низким, за интересующий нас временной интервал. Для каждого пикселя вычислить вегетационный индекс NDVI.*

Полученный таким образом ряд значений может быть использован для дальнейшего анализа различными методами, предназначенными для анализа временных рядов [13, 14]. В случае если изображение участка содержится в нескольких пикселях, может быть оценено распределение значений (рис. 1).

Предположим, что каждый пиксель содержится в географической базе данных, где определены размер и контур участка поверхности, изображение которого содержится в пикселе, дата регистрации изображения, и измерения в различных полях таблицы. В реляционных базах данных стандартизовано представление географической информации и отметок времени. В то же время, стандарты представления физических величин отсутствуют, поэтому информация, необходимая для интерпретации должна быть также представлена в этой базе данных.

В отношении нашего примера необходимо для каждого пикселя данных дистанционного зондирования иметь возможность определить:

1) в каких полях таблицы содержатся измерения, выполненные в диапазонах длин волн, соответствующих максимуму поглощения растительностью в красной части видимой части спектра и максимуму отражения в ближнем инфракрасном диапазоне,
2) в каких единицах эти измерения выполнены,
3) в каких полях таблицы содержится информация о состоянии атмосферы и как её интерпретировать,
4) какова точность выполненных измерений.

Последнее значение может быть необходимо для оценки доверительных интервалов значений вычисляемых функций и для принятия решений на основании полученных данных.

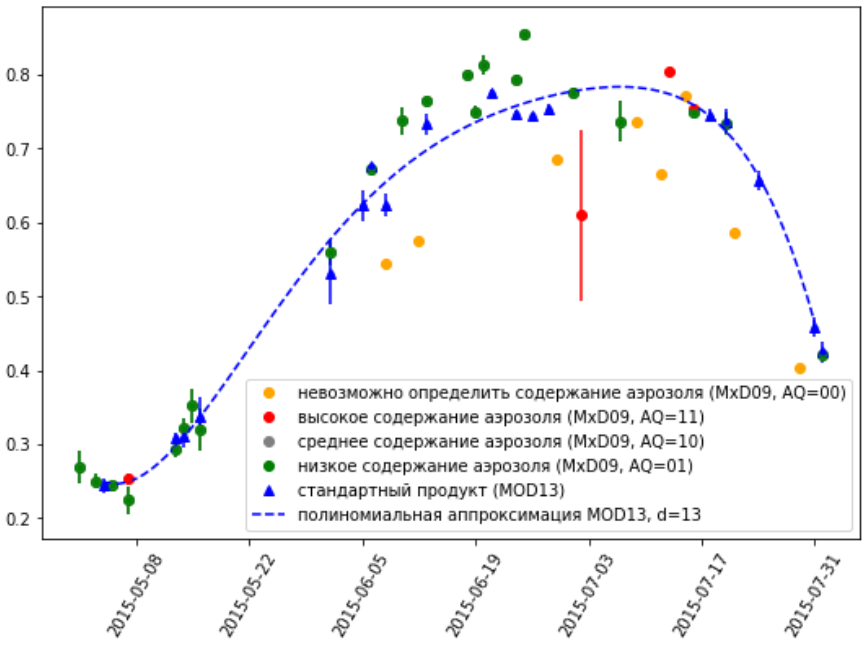

Рис. 1 – Значения NDVI, рассчитанные по данным, выбранным из нескольких таблиц

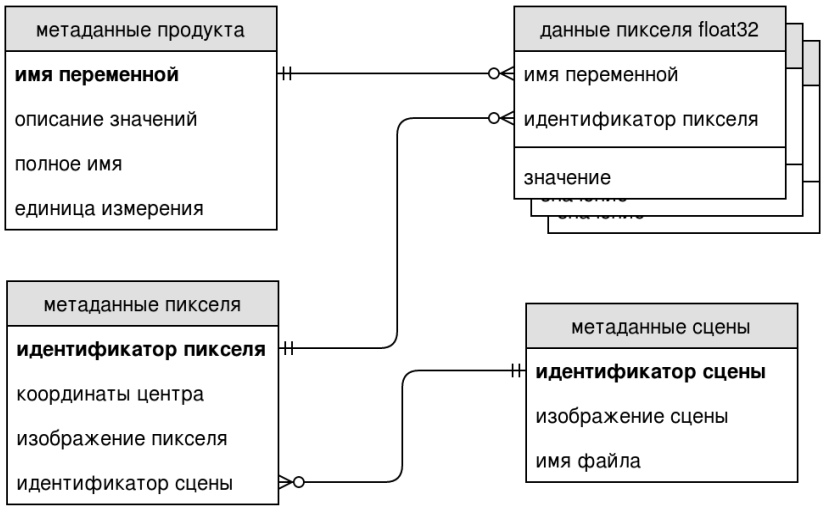

Рис. 2. ER-диаграмма для схемы данных, представляющих данные и метаданные.

Соответствующая реляционная схема может содержать таблицы для всех уровней метаданных, а также отдельные таблицы для различных скалярных типов (Рис. 2). Фактическая реализация может отклоняться от этой схемы из практических изображений, либо исходя из заранее ограниченного перечня представленных переменных.

В этом примере предпринимается попытка выстроить единую схему, позволяющую выбирать данные по их атрибутам, т.е. не изучая заранее структуру каждого продукта в отдельности. При этом следует учитывать, что большая гибкость схемы приведет к усложнению запросов к ней. Это противоречие может быть разрешено разработкой языка запросов высокого уровня, скрывающего сложность схемы данных.

Выводы. Предложенный подход позволяет дополнить представление данных в реляционной схеме представлением всех используемых при обработке данных дистанционного зондирования уровней метаданных. Таким образом, все данные, представленные в блоках, содержащихся в файлах форматов HDF или netCDF, отображены в единую реляционную схему. Такая схема также может уменьшить степень зависимости конкретных запросов к данным от специфики используемых информационных продуктов, что позволит без изменений выполнять вычисления с измерениями, на основе данных, полученных с помощью различных сенсоров.

С возрастанием числа спутниковых платформ и приборов необходимость систематического подхода к хранению данных для обеспечения их эффективного поиска становится всё более актуальной. На решение этой задачи направлен целый ряд проектов, начиная от Global Change Master Directory [15] и заканчивая проектами, основанными на семантических моделях представления информации [16–18].

Работа выполнена при финансовой поддержке программ ФНИ (проекты № 0316-2016-0002, № 0330-2016-0018) программы фундаментальных исследований президиума РАН и гранта Президента РФ для государственной поддержки ведущих научных школ (грант № НШ-7214.2016.9).

ЛИТЕРАТУРА

[1] Kuenzer C., Dech S., Wagner W. Remote sensing time series revealing land surface dynamics: Status quo and the pathway ahead // Remote Sensing Time Series. Springer International Publishing, 2015. P. 1-24.

[2] Шокин Ю.И., Добрецов Н.Н., Мамаш Е.А. и др. Информационная система приема, обработки и доступа к спутниковым данным и ее применение для решения задач мониторинга окружающей среды // Вычислительные технологии. 2015. Т. 20, № 5. С. 157-174.

[3] Goucher G., Mathews G.A. Comprehensive look at CDF // National Space Science Data Center Publication, 1994. N 94-07.
[4] Rew R., Davis G. NetCDF: an interface for scientific data access // IEEE computer graphics and applications. 1990. Vol. 10, N 4. P. 76-82.
[5] Folk M., McGrath R.E., Yeager N. HDF: an update and future directions // Proceedings of International Geoscience and Remote Sensing Symposium. 1999. Vol. 1. P. 273-275.
[6] Rew R.K., Hartnett E.J. Merging NetCDF and HDF5 // Proceedings of the 21st International Conference on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology. 2004. P. 1-11.
[7] Cornillon P., Gallagher J., Sgouros T. OPeNDAP: Accessing data in a distributed, heterogeneous environment // Data Science Journal. 2003. Vol. 2. P. 164-174.
[8] Baumann P., Dehmel A., Furtado P. et al. The multidimensional database system RasDaMan // ACM Sigmod Record. 1998. Vol. 27, N 2. P. 575-577.
[9] Fiore S., Palazzo C., D'Anca A. et al. A Big Data Analytics Framework for Scientific Data Management // Proceedings of 2013 IEEE International Conference on Big Data. 2013. P. 1-8.
[10] Lewis A. et al. The Australian Geoscience Data Cube-Foundations and lessons learned // Remote Sensing of Environment. 2017. Vol. 202, N 1. P. 276-292.
[11] Klein L. Taaheri A. HDF-EOS5 Data Model, File Format and Library. NASA ESDS-RFC-008.v1.1. 2016. P. 1-52.
[12] Vermote E.F., Kotchenova S.Y., Ray J.P. MODIS Surface Reflectance User's Guide. MODIS Land Surface Reflectance Science Computing Facility. Version 1.3. 2011. P. 1-40.
[13] Coppin P., Jonckheere I., Nackaerts K. et al. Digital Change Detection Methods in Ecosystem Monitoring: a Review // International Journal of Remote Sensing. 2004. Vol. 25, N 9. P. 1565-1596.
[14] Hussain M., Chen D., Cheng A. et al. Change detection from remotely sensed images: from pixel-based to object-based approaches // ISPRS Journal of Photogrammetry and Remote Sensing. 2013. Vol. 80. P. 91-106.
[15] Gordov E.P., Okladnikov I.G., Titov A.G., Fazliev A.Z. Some Aspects of Development of Virtual Research Environment for Analysis of Climate Change Consequences // Data Analytics and Management in Data Intensive Domains. 2016. P. 291-297.
[16] Bart A.A., Fazliev A.Z., Privezentsev A.I. et al. Ontological description of meteorological and climate data collections // Collection of Scientific Paperes of the XIX International Conference DAMDID Data Analytics and Management in Data Intensive Domains. 2017. P. 340-346.
[17] Major G.R. Beyond Bibliography: A Dynamic Approach to the Cataloging of Multidisciplinary Environmental Data for Global Change Research // Online Ecological and Environmental Data. 2014, P. 21-36.
[18] Visser U., Stuckenschmidt H., Wache H., Vögele T. Using environmental information efficiently: Sharing data and knowledge from heterogeneous sources // Environmental information systems in industry and public administration. 2001. P. 41-73.