Narzędzia

CTAN – plany*

Robin Fairbairns, Jim Hefferon,
Rainer Schöpf, Joachim Schrod,
Graham Williams i Reinhard Zierke

Abstract

The readers of TUGboat likely know the Comprehensive TeX Archive Network as a great pile of TeX stuff. That is, it is full of TeX materials and it is great, but it is also perhaps a pile — a bit of a mess.

We will sketch some plans for improving CTAN. As part of that, we will outline its architecture, history, and some issues.

Streszczenie

Czytelnicy TUGboat-a zapewne znają CTAN, *Comprehensive TeX Archive Network*, jako wielką stertę zasobów TeX-owych. Znaczy to, że CTAN jest wypełniony zasobami TeX-owymi i jest wielki, ale jest też pewnie stertą – bałaganem.

Naszkicujemy plany ulepszenia CTAN-u. Jako jedną z części tego planu opiszemy jego architekturę, historię i niektóre problemy.

1. Wstęp

Zacznijmy od początku: CTAN jest internetowym archiwum zasobów związanych z TeX-em, udostępnionych do użytku publicznego. Mamy obecnie pięć gigabajtów tych zasobów. Codziennie około dziesięciu tysięcy odwiedzających pobiera wielką liczbę plików, zaś inni dostarczają nowe zasoby.

Bardzo się staramy stosować dosłownie do przymiotnika *comprehensive*[1]. Przechowujemy wszystko, począwszy od pakietów makr LaTeX-owych po kompletne dystrybucje takie jak MikTeX i TeTeX.

2. Stan obecny

CTAN nie jest pojedynczym serwerem, ale zbiorem serwerów. Trzy z nich tworzą jego jądro i aktywnie zarządzają zasobami, na przykład przez instalowanie nowych lub zaktualizowanych pakietów.

- `dante.de.org` w Niemczech jest sponsorowany przez Dante, niemiecką grupę użytkowników TeX-a, i administrowany przez Rainera Schöpfa oraz Reinharda Zierke.
- `cam.ctan.org` jest sponsorowany przez UK-TUG w Anglii i utrzymywany w ruchu przez Robina Fairbainsa.
- `tug.ctan.org` w USA jest sponsorowany przez TUG a utrzymywany przez Jima Hefferona.

W uzyskaniu jednolitości zarządzania pomaga nam aktywna wymiana wiadomości na liście dyskusyjnej. Do synchronizowania zawartości stosujemy własne, specjalizowane skrypty.

Oprócz wymienionych wyżej głównych serwerów, pomaga nam około siedemdziesięciu pięciu lokalizacji na całym świecie poprzez *mirrorowanie* – synchronizują się one codziennie z jednym z serwerów centralnych a następnie udostępniają wynik tej synchronizacji publicznie. Daje to użytkownikom większą swobodę wyboru i zmniejsza ruch sieciowy. Zachęcamy do używania mirrorów[2].

3. Przeszłość

Przed CTAN-em istniały serwery udostępniające materiały TeX-owe, ale nie istniała jedna, autorytatywna kolekcja tych zasobów. W czasie dyskusji panelowej zorganizowanej przez Joachima Schroda na konferencji EuroTeX w 1991 r. narodziła się idea zjednoczenia osobnych kolekcji. Joachim zaangażował się w jej realizację ponieważ zarządzał wtedy, przy pomocy przystosowanego przez siebie narzędzia `mirror.pl`, jednym z największych w owym czasie serwerów FTP w Niemczech.

CTAN został zbudowany w 1992 r. przez Rainera Schöpfa i Joachima Schroda w Niemczech, Sebastiana Rahtza w Wielkiej Brytanii oraz George Greenwada w USA (nazwę wymyślił George). Struktura została zdefiniowana na początku 1992 r. – główną pracę wykonał Sebastian – i zsynchronizowana na początku 1993 r. TUG ustanowił ramy organizacyjne – *Technical Working Group* – do prac nad tym zadaniem. Oficjalna premiera CTAN odbyła się podczas konferencji EuroTeX w Aston, w 1993 r.

---

* Tekst jest tłumaczeniem artykułu zamieszczonego w numerze 24/2 TUGboata (2003). Dziękujemy za zgodę na przedruk redaktorom TUGboata i autorom. Tłumaczenia z języka angielskiego dokonał Jerzy Ludwichowski (*red.*).

[1] *comprehensive* – pełny (*przyp. tłumacza*).

[2] Zobacz `http//www.dante.de/mirmon/` lub `ftp://tug.ctan.org/tex-archive/README.mirrors`.

Kiedy tworzono CTAN, główną metodą dostępu do jego plików poprzez sieć był FTP. System został więc zbudowany przy założeniu, że odwiedzający będą pobierać zasoby w taki właśnie sposób (być może założono również, że będą to użytkownicy doświadczeni). W roku 1999 na serwerze TUG podjęto próbę zastosowania bardziej rozbudowanego dostępu poprzez WWW, ale pomysł nie został podchwycony przez pozostałe dwa serwery ze względu na jego słabości realizacyjne.

4. Problemy

Nikt nie lubi narzekających, ale aby opisać nasze plany musimy opisać problemy, których one dotyczą. Są to problemy zarówno z samą kolekcją jak również z jej zarządzaniem.

Jednym z problemów jest to, że kolekcja jest duża. Wyrosła ona ze swojej dotychczasowej struktury, która wymaga zaktualizowania. Większość członków społeczności TEX-owej pewnie zna z własnego doświadczenia sytuację niemożności rozwiązania jakiegoś problemu TEX-owego, po której nastąpiło odkrycie, że rozwiązania istnieje na CTAN-ie. Tak więc stwierdziliśmy, że wraz z naszym rozwojem informacja dostępna dla użytkowników archiwum nie rosła wystarczająco szybko aby byli oni w stanie znajdować to, czego potrzebują. Sytuację złagodziły metadane[3] zebrane przez Grahama Williamsa w jego Katalogu (*Catalogue*)[4]. Mimo to musimy stać się informacyjnie bogatsi.

Dawny postulat odnośnie kolekcji[5] dotyczy utrzymywania przez CTAN historii pakietów w taki sposób, aby użytkownicy mogli kompilować dokumenty, które zależą od ich starszych wersji. W chwili obecnej, jeśli autor dostarczy aktualizację, nadpisujemy nią stary zasób.

Inny problem w kategorii *nie-pomyśleliśmy-że-urośniemy* (lub *się-zestarzejemy*) dotyczy mirrorów. Dla użytkownika często najlepszym sposobem uzyskania pakietu z CTAN bez pominięcia jakiegoś pliku jest załadowanie całego folderu. Dla ułatwienia takiego podejścia, serwery centralne umożliwiają tworzenie skompresowanych paczek plików `.zip` i `.tar.gz`.[6] Interfejs WWW na `http://www.ctan.org` używa tej właściwości i należy zapewnić tę funkcjonalność we wszystkich przyszłych interfejsach. To jednak nie działa w przypadku *mirrorów*. Aby przekierować użytkownika, który chce skompresowaną paczkę, system musi wiedzieć, które z mirrorów poprawnie wykonują kompresję w locie. Napisaliśmy więc skrypty, które to sprawdzają. Kiedy je uruchomiliśmy okazało się, że ani jeden mirror nie wykonał zarówno plików `.zip` jak i `.tar.gz` bez błędu. W efekcie większość zasobów pobieranych przez użytkowników pochodzi z trzech serwerów jądra.

Abyśmy mogli dostarczać zasoby, musimy je najpierw otrzymać od społeczności. Jesteśmy zaniepokojeni zaznaczającą się tendencją polegającą na tym, że niektórzy autorzy nie przekazują swojego dorobku do archiwum, ale udostępniają je poprzez osobiste serwery WWW. Ta tendencja jest niedobra, ponieważ cofa nas w czasy sprzed CTAN-u, czasy zasobów rozproszonych, zasobów, które mogą zaginąć; tak więc CTAN mógłby stać się nie-pełny (*not-comprehensive*). Nawet wtedy gdy wiemy, że praca danego autora istnieje, mamy kłopoty z włączeniem takiego materiału ponieważ protokół HTTP nie ułatwia jego pobierania do naszych zbiorów[7]. Oczywiście, musimy przyjąć świat takim jakim jest, ale problem istnieje.

Ostatnim problemem kolekcji, który wymienimy, jest to, że wcześni twórcy, łącznie z Knuthem, zakładali, że użytkownicy będą stosunkowo zaawansowani: będą posiadali podstawową wiedzę typograficzną i będą potrzebowali jedynie minimalnego wsparcia (będą pisali własna makra). Okazało się, że jest inaczej. Większość użytkowników TEX-a chce dystrybucji, które będą zawierały LATEX-a i inne pakiety już przygotowane do użytku. Jeśli potrzebują czegoś więcej, to chcieliby, aby dystrybucja posiadała moduł, który współpracując z CTAN-em przygotuje pakiet do użycia. Tak więc kluczowym celem jest przygotowanie systemu, który – we współpracy z dystrybucjami – spełni takie oczekiwania.

Dalej opiszemy niektóre problemy związane z administrowaniem archiwum. Użytkownicy nie

[3] Dane o danych, czyli dane o pakietach.

[4] `http://ctan.org/tex-archive/info/Catalogue`.

[5] Pochodzi on w szczególności od Nelsona Beebe.

[6] Np. odwiedzając `ftp://ftp.ctan.org/tex-archive/macros/latex/contrib/shadethm.zip` uzyska się cały folder `shadethm` jako archiwum zip-owe. Komenda klienta FTP `get /tex-archive/macros/latex/contrib/shadethm.zip` wywoła taki sam skutek.

[7] W skali, w jakiej pracujemy, nie wystarczy użycie programu takiego jak np. `wget`.

widzą ich bezpośrednio, ale mają one wpływ na ich percepcję archiwum.

Pierwszy wynika z tego, że posiadamy do swojej dyspozycji niewielkie środki. Maszyny zostały zakupione ze środków grup użytkowników, a krytyczne dla ich działania łącza sieciowe są sponsorowane przez instytucje, w których pracują administratorzy. Administratorzy są bardzo od siebie oddaleni – niektórzy członkowie listy dyskusyjnej nigdy nie spotkali żadnego innego członka tej samej listy, co spowalnia postęp i zwiększa niebezpieczeństwo nieporozumień komunikacyjnych. Wszyscy administratorzy są wolontariuszami i mogą poświęcać czas na zarządzanie CTAN-em tylko wtedy, gdy pozwalają im na to inne obowiązki.

Koszty administratorów w kategoriach czasu są ważne, ponieważ powodowały opóźnienia naszego rozwoju. W szczególności, nowo dostarczonymi zasobami należy się zająć możliwie najszybciej. Aby pomóc czytelnikowi uświadomić sobie ten problem – i aby trochę ponarzekać – rozważmy pojawienie się nowego pakietu. Administrator maszyny pobiera go, rozpakowuje, sprawdza zasady licencjonowania i decyduje, gdzie pakiet ma zostać umieszczony. Sprawdza, czy w pakiecie jest plik README i czy jest dokumentacja w formacie PDF, która używa fontów Type 1. Często prowadzi to do wymiany korespondencji z autorem albo na liście dyskusyjnej CTAN, co wprowadza opóźnienia jedno- lub wielodniowe. Po tych wszystkich uzgodnieniach administrator sięga do naszego lokalnego skryptu instalacyjnego i kopiuje zasoby do obszaru dostępnego publicznie oraz inicjalizuje proces *mirrorowania*. Następnie zawiadamia listę ogłoszeń CTAN (a tym samym `comp.text.tex`). W końcu aktualizuje metadane w Katalogu (*Catalogue*) i dodaje je do drzewa CVS. Średni nakład pracy na pakiet wynosi około pół godziny.

Pomocą może być zwiększenie liczby administratorów. Trzeba jednak zdawać sobie sprawę z tego, że oprócz koniecznej znajomości TEX-a, umiejętności administrowania systemami i znajomości reguł rządzących CTAN-em w ostatnich latach przychodzi nam działać w coraz bardziej złożonym świecie przetwarzania informacji. Jednym z wielkich problemów ostatnich lat stały się zagadnienia licencjonowania. Potrzebujemy współpracowników, ale potrzebujemy również metody włączania ich do pracy w taki sposób, aby mogli się wdrażać stopniowo.

5. Plany

Załóżmy, że kolega przyśle ci artykuł wymagający do kompilacji pakietu, który nie jest zawarty w Twojej instalacji TEX-a. W aktualnym stanie rzeczy odwiedziłbyś CTAN, odnalazł pakiet i zainstalował go. Wyobraź sobie, że zamiast tego Twoja dystrybucja TEX-a odnajdzie pakiet, zainstaluje go i wykona dalsze czynności związane z kompilacją artykułu bez żadnej ingerencji z Twojej strony. Technologia, która umożliwia tego rodzaju negocjacje pomiędzy komputerem użytkownika i CTAN-em nazywa się usługami WWW (*Web Services*)[8]. Naszym najważniejszym zadaniem jest zbudowanie – w porozumieniu z istniejącymi dystrybucjami – dobrej oferty usług WWW dla CTAN.

Jednym z kroków w tym kierunku jak również jednym z bieżących celów jest lepsze zorganizowanie naszego stanu posiadania. Na przykład połączyliśmy podkatalogi `supported` i `other` katalogu `/macros/latex/contrib` oraz planujemy połączenie katalogów `/info` i `/help`. Większym zadaniem jest zdekomponowanie naszego stanu posiadania w pakiety, z których każdy rezydowałby w swoim własnym katalogu (nigdy więcej `misc`). Jest to naturalny sposób udzielenia odpowiedzi na skierowane do usług WWW pytanie: „Jaka jest najnowsza wersja pliku `f` w pakiecie `p`?".

Jako wsparcie usług WWW oraz pomoc dla odwiedzających w uzyskiwaniu większej ilości informacji musimy wprowadzić więcej informacji do Katalogu (*Catalogue*). Musimy (1) rozszerzyć ilość informacji w kategoriach już w nim obecnych i (2) rozszerzyć go o nowe kategorie informacji.

Część (1) idzie w kierunku dostarczenia łatwego sposobu aktualizowania tych metadanych przez WWW, na przykład wtedy, gdy autor dostarcza nowy lub modyfikuje istniejący pakiet. Jako dodatkowy zysk pojawia się możliwość włączania nowych osób do pomocy przy CTAN-ie. Sensownym współdziałaniem przy jego utrzymywaniu byłoby aktualizowanie metadanych i wprowadzanie ich do drzewa CVS bez konieczności wykonywania prac związanych z administrowaniem serwerem CTAN.

Przykładem z obszaru (2) jest konieczność utrzymywania słów kluczowych (*keywords*) w taki sposób, aby przy ich pomocy użytkownicy mogli

[8] Zgrubna definicja: serwer WWW odpowiada na żądania wykraczające poza proste dostarczanie stron.

szukać pakietów (takie poszukiwanie mogłoby następować na stronie WWW CTAN lub na stacji roboczej przy użyciu usług WWW). Od pewnego już czasu dyskutowaliśmy leżący u podstaw tego model metadanych i – jako pierwszy krok w tym kierunku – przenieśliśmy ostatnio Katalog *Catalogue* do drzewa CVS[9].

Cała informacja powinna znajdować się w bazie danych. To pasuje do naszych planów pod wieloma względami, ponieważ od czasu gdy CTAN powstał jako archiwum FTP, technologie WWW zmieniły praktycznie wszystko z takim skutkiem, że musimy poprawić system tak, aby był wspierany technologią bazodanową. Opracowany interfejs powinien być jednolity dla wszystkich serwerów jądra.

Interfejs ten mógłby pozwalać użytkownikom na odnajdywanie pakietów na różne sposoby (pierwszy zaproponował to Sebastian w jednym ze swoich komentarzy). Aktualnie użytkownicy mogą przeglądać katalogi FTP, mogą przeglądać listę wszystkich plików, mogą wykonać prymitywne przeszukania tekstowe Katalogu (*Catalogue*), mogą też przeszukać nasze zasoby przy użyciu standardowej wyszukiwarki WWW i – jak już wspomnieliśmy – chcemy dodać przeszukiwanie według słów kluczowych. Ponadto chcielibyśmy dodać poszukiwanie pakietów według funkcjonalności: użytkownik pracujący nad nagłówkami mógłby „przeklikać się" przez gałąź wyborów takich jak `Top > LaTeX > Page layout > Headers and footers`[10].

Jedną z własności, którą powinna posiadać współczesna witryna jest zdolność do przeszukiwania dokumentacji. W chwili obecnej wiele pakietów nie posiada dokumentacji, lub posiada ją w formacie, który nie nadaje się do prezentacji użytkownikowi jako wynik wyszukiwania (np., jeśli wynikiem wyszukania jest łącze do pliku `.dtx`, to kliknięcie w nie najprawdopodobniej niewiele pomoże). Rozpoczęliśmy wymuszanie na autorach pakietów dostarczania dokumentacji w formacie PDF, który jest jedynym formatem łączącym szeroką dostępność i wysoką jakość typograficzną.

Pośród wyżej wymienionych problemów znalazły się utrzymywanie historii pakietów i zagadnienie dostarczania przez mirrory paczek `.zip` i `.tar.gz`. Wydaje nam się, że można te problemy rozwiązać łącznie, archiwizując każdą wersję pakietu jako paczkę – mielibyśmy gotowe paczki, których nie musiałyby tworzyć *mirrory*.

Musimy przekonać autorów aby dostarczali nam swoje zasoby. W przeszłości namawialiśmy autorów aby tak robili[11], ale mogliby tu pomóc wolontariusze, którzy znalazłszy zasoby mogliby uprzejmie namówić autora.

I w końcu: myślimy ciągle o zadaniach (*work flow*) stojących przed administratorami serwerów. Prowadzono chaotyczne rozmowy o GUI (*Graphical User Interface*, interfejs okienkowy) administratora, ale problemem jest istnienie tak wielu wyjątków i przypadków specjalnych, że często nie widzimy innego sposobu niż ręczne wykonywanie tych prac.

6. Rokowania

Planujemy, aby CTAN stał się bardziej „zrozumiałym" (*more of a "Comprehensible"*) sieciowym archiwum TeX-owym. Plany te są dyskutowane od dwóch lub trzech lat.

Bardzo w tym pomogło Dante przez sfinansowanie udziału kluczowych osób w spotkaniach twarzą w twarz w ostatnich 2 latach w Bremen i Darmstadt. Mamy przyjemność powiedzieć, że opiekujące się nami osoby z Dante wykraczały znacznie poza zwyczajową grzeczność i gościnność: w szczególności dzięki Volkerowi i Klausowi cała sprawa posunęła się bardzo mocno do przodu.

Aktualnie, format Katalogu (*Catalogue*) został dopasowany w taki sposób, aby umożliwić jego rozwój (oprócz tego, przeniesienie go do drzewa CVS umożliwiło równoległe aktualizacje), zostały zbudowane struktury bazodanowe i istnieje wersja beta kodu pozwalającego na edytowanie metadanych przez WWW, istnieją również inne części nowego systemu WWW. Musimy również dostarczyć dane statyczne: zawartość stron WWW, słowa kluczowe, kategorie funkcjonalne (*by-function categories*) itp. I na koniec: potrzebujemy więcej danych do bazy danych o pakietach.

W skrócie: postęp jest denerwująco wolny, ale *jest*.

◇ Robin Fairbairns et. al
ctan@dante.de

[9] `http://texcatalogue.sarovar.org`.

[10] Autorzy artykułu nie wspomnieli tu o internacjonalizacji – „przeklikanie się" przez tę gałąź wyborów wymaga pewnej znajomości języka angielskiego (*przyp. tłumacza*).

[11] Prosimy – rozważ, czy możesz podzielić się czymś, co inni mogliby uznać za pożyteczne!